Quality Statistical Review Checklist of Investigational Device Exemption (IDE) Submissions for Therapeutic and Aesthetic Medical Devices

Gregory Campbell, Lilly Yue, Yonghong Gao, Martin Ho, Heng Li, Vandana Mukhi, Laura Thompson, Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993.

An Investigational Device Exemption (IDE) is a submission to the FDA that, if approved, allows a sponsor, who could be an investigator or a device company or some other entity, to begin a clinical investigational study in the United States on humans when that study could pose a potential significant risk. IDE approval by FDA is required before such significant risk studies can be undertaken in the United States. The time limit for FDA to take an action such as approval, approval with conditions or disapproval is 30 days. If not such action is taken then the IDE is automatically approved. If a study does not pose significant risk, is not life-saving nor life-threatening, then no IDE would be required. So, for example, for many studies of *in vitro* diagnostic products based on human specimens such as blood or urine and for which no medical action is undertaken as a result of the outcome of the investigational test then no IDE would be required. Of course, institutions in the US at which investigational studies on humans are undertaken require also study review by their Institutional Review Boards (IRBs) as well.

This article explains the internal quality review procedure for the statistical review of such IDE submissions to FDA. As such it is really more than a checklist but rather a guide to insure a complete statistical review by the statistical reviewer of the IDE.

The FDA Safety and Innovation Act (FDASIA) was signed into law on July 9, 2012, and has a number of provisions. One provision amends Section 520(g)(4)(C) of the Food Drug and Cosmetic Act to indicate that FDA shall not disapprove an IDE because:

- *the investigation may not support a substantial equivalence or de novo classification determination or approval of a device;*
- *the investigation may not meet a requirement, including a data requirement, relating to the approval or clearance of a device; or*
- *an additional or different investigation may be necessary to support clearance or approval of the device.*

In light of the passage of this law, this internal quality review procedure for the statistical review of IDEs is for the purpose of addressing the study design in the proposed IDE and may not by itself result in the disapproval of the IDE but could point out serious study design issues that could threaten the scientific validity of the investigation and its value in providing valid scientific information for a marketing application such as a PreMarket Notification (called a 510(k) submission) or a PreMarket Approval (PMA) application.

This article pertains to the review of what FDA calls pivotal clinical investigations. The recent draft guidance on Pivotal Clinical Study Design identified three separate stages for clinical studies for medical devices (FDA, 2011). They are:

- Exploratory Stage. This includes the first-in-human, feasibility and pilot studies that could result in iterative learning about the design of the product and its continued refinement.
- Pivotal Stage. This includes all definitive studies to support the safety and effectiveness evaluation of the medical device for its intended use. This stage would correspond to a Phase III study for an investigational drug, sometimes also called a confirmatory study.
- PostMarket Stage. This encompasses studies intended to better understand the long-term effectiveness and safety of the device, including rare adverse events.

It is important to note that, unlike the four phases of drug development, it may not be necessary for every investigational product to go through every stage. In particular, if there are extensive non-clinical studies, including bench testing and animal studies, it may be possible to forego the exploratory clinical stage. However, it is often extremely valuable not to rush through this stage in an effort to get to the pivotal stage.

The other important note is that this article is limited to therapeutic and aesthetic devices under IDE review. The draft guidance identifies three different kinds of devices:

- Therapeutic devices. These are devices that are intended to treat a specific condition or disease.
- Aesthetic devices. These devices provide a desired change in a subject's appearance through physical modification or the structure of the body. An example device would be a wrinkle filler.
- Diagnostic devices. These devices provide information when used alone or in the context of other information to help asses a subject's condition. It could be to screen or detect disease, to monitor a condition or to provide a risk assessment, for example.

The purpose of this article is provide internal guidance to the statistical reviewer but also to provide transparency to the device industry as to what an FDA statistical reviewer looks for in the statistical review of an IDE. The purpose of this current effort is to facilitate high-quality, consistent statistical reviews of IDEs by providing items to consider when reviewing an IDE. The external purpose is to share it with industry through the presentation at the Joint Statistical Meetings and this accompanying article in order to increase the consistency, transparency and predictability of the statistical review process for industry and hopefully result in an increase in the quality of IDE submissions. In particular, this is an effort to make it clear to sponsors what the FDA statistical reviewer is looking for in the assessment of the IDE in terms of what would constitute a scientifically valid study that would support a marketing application. There is also in this proceedings a companion article concerning the statistical review of diagnostic IDEs (Vishnuvajjala et al, 2013).

These two articles on IDEs in this proceedings parallel a similar effort concerning PMAs, with the 2007 JSM Proceedings having two PMA checklists, one for non-diagnostic PMAs (Yue, 2007) and the other for diagnostic ones (Vishnuvajjala, 2007).

Topics in the accompanying checklist are now addressed in the next five sections.

I. Introduction/Background

In this section of the statistical checklist, the statistical reviewer will assess whether the purpose of the study is clearly specified, i.e., is the scientific rationale for why the study is being performed provided and will it provide data that potentially could support a premarketing submission, either a Pre-Market Approval (PMA) or a Pre-Market Notification (510(k)) application. The reviewer will look for a clear indication-for-use statement, providing a general description of the disease or condition the device will treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended. It is also useful to provide all relevant background clinical information that may be available for the investigational device for its proposed intended use. This may include relevant results from all, previous or on-going studies, feasibility and/or pivotal studies and studies conducted in US or outside of US (OUS), for the investigational device for its intended use. Lastly, has the regulatory history of the device been adequately and completely reported in the IDE submission?

II. Study Design

In the Study Design section of the checklist, the statistical reviewer will address the question of whether the IDE submission contains adequate details concerning the design of the study, including, but not limited to, the type of study design, number of treatment arms, description of each intervention, randomization scheme, blinding scheme, choice of appropriate control(s), screening criteria, follow-up schedule, missing data prevention plan, number of US/OUS sites, and data quality monitoring plan.

In general, double-blinded (double-masked), randomized, controlled, multi-center clinical trials (RCTs) provide the strongest level of scientific evidence and are considered the “gold standard”, as these designs tend to minimize bias. But such studies may be impractical, unethical or infeasible in certain situations. For example, when a RCT is conducted with an active or placebo control, it may not be possible to blind (mask) the subjects or the investigators, or even a third-party evaluator to the intervention (treatment) assignment. In such cases, bias may be introduced which could lead to an incorrect determination of safety and effectiveness. The reviewer will then look for any proposed strategies that can help eliminate or minimize bias and also see if issues of group comparability have been addressed.

Sometimes clinical equipoise may not exist to randomize subjects to two treatment groups. In such situations, a non-randomized study with concurrent or nonconcurrent/historical control group may be considered. But such a study design may result in two treatment groups that may not be comparable due to imbalance in the covariates. Propensity score (PS) methods provide an analytical tool to adjust for imbalances in observed covariates; however caution is needed as it may be inappropriate when sample size is small (Yue (2007), Yue (2012), Levenson and Yue (2013)). Also, such methods can only help adjust for observed covariates and not for unobserved ones. This is an important consideration when applying propensity score methods to an historical control study where significant baseline data may not be available. If applicable, it is useful to pre-specify complete details of propensity score method in the Statistical Analysis Plan (SAP) section of the protocol. The logistics of how PS methods will be implemented, either without any access to outcome data or conducted by an independent statistician who has no access to such data, would also need to be described in detail.

For a one-arm study design, the investigational device results on primary safety and effectiveness outcomes are compared to a fixed value. This can either be a performance goal (PG) or objective performance criteria (OPC). Such one-arm studies should be considered as a last option when considering designing a pivotal clinical study since the bias in such studies is unknown and can be quite large. With this study design, one can only claim if the pre-specified PG or OPC has been met or not. Since there is no control group involved in such studies, comparison to an OPC or PG cannot demonstrate either superiority or non-inferiority and thus such claims cannot be made. Also, achievement of (or failure to achieve) a PG does not necessarily lead to immediate acceptance (or rejection) of the study results. In some cases, the study results need to be explored more qualitatively if they are mixed or if unusual signals within the results are found. Thus, such clinical study designs should be used with great care.

III. Primary and Secondary Clinical Endpoints

The most fundamental element of designing a pivotal clinical study is selection of an appropriate clinical endpoint, i.e. selecting an endpoint or endpoints that are objective, reflect the intended use of the device, represent a clinically meaningful outcome for patients, are easily ascertainable and can be measured with minimal bias. The reviewer will look to see that the IDE includes clear definitions of primary safety and effectiveness endpoint(s) and how they will be evaluated to demonstrate device performance. If multiple primary effectiveness (or safety) endpoints are specified, did the protocol provide a scientific rationale and explain the role and relative importance of each endpoint?

It is recommended that verbal statements and mathematical expressions of the statistical hypothesis associated with each primary endpoint be included in the study protocol. The reviewer will further examine the IDE to see whether study success/failure criterion with respect to each of multiple primary endpoints has been pre-specified, along with appropriate statistical approaches to handling multiplicity issues and controlling overall Type I error rate. When multiple secondary endpoints are selected with potential additional claims in mind, the reviewer will examine the protocol to see whether appropriate statistical methods to analyze data and interpret results for these secondary are pre-specified. For two-arm studies, if non-inferiority comparisons are being made, there will be a check to see if a clinical justification of the choice of non-inferiority margin is included. Similarly, for one-arm studies, the justification of a performance goal (PG) is essential.

IV. Sample Size

An IDE typically contains a justification of the number of subjects of the study in a section called sample size determination. In this section of the IDE, the reviewer will look for to see that a statistical justification is provided the number of subjects the study plans to enroll. An adequate justification of the proposed sample size identifies the primary effectiveness endpoint or co-primary endpoints (and the safety endpoints, if appropriate). In the case of a study that involves statistical hypothesis testing, it is expected that the IDE also includes the null and alternative hypotheses (both in words and in mathematical expression), This may include details regarding the test statistics used to calculate sample size, Type I error and Type II error rates assumed, expected effect size, assumptions made on parameters, variability assumptions and correlation assumed among endpoints. The reviewer will check to see if the calculation is done correctly. In addition, the statistical reviewer will check to see if the IDE contains a justification of the assumptions on the parameters, including the variability of endpoints, the treatment effect size, and the correlation between endpoints, if appropriate. Furthermore, the reviewer will look for any details regarding the expected rate of withdrawals, protocol violations, and missing data that may be taken into consideration in the sample size calculation. Finally, given the proposed sample size, the possibility of observing dissonance between statistical significance and clinically significance may also be included.

For adaptive designs such as Bayesian adaptive design or group sequential design, a study’s sample size is not usually fixed in advance and thus additional characterization will be needed. Details about an IDE’s criteria for adaptive designs will be discussed in the latter part of this paper and also covered in a Supplement to the Checklist.

V. Statistical Analysis Plan

The Statistical analysis plan (SAP) of an IDE provides important details about how to evaluate the safety and effectiveness of a treatment based on the study results and what assumptions have been made at the planning stage of a pivotal study. Despite the wide variety of pivotal study designs for a diverse array of medical devices and therapeutic areas reviewed by FDA, the SAP components discussed in this section are considered essential in establishing sufficient level of evidence that leads to a successful pre-market application.

It is a good practice to provide a list of baseline and time-dependent covariates, and outcome variables to be measured on the study subjects (such as subject demographics, subject's baseline severity of illness, or other important subgroup characteristics) and describe the summary statistics of these variables in a SAP. Based on these baseline characteristics, a plan to assess the balance of these baseline characteristics between different study groups will provide critical insight when interpreting the study results. Moreover, for medical device trials, it is important to demonstrate the consistency of observed treatment benefit across surgeons, sites, geographic regions, and other potentially significant subgroups in so-called data poolability analysis. If the study result is shown to be not poolable, a prospectively specified alternative analysis plan will be valuable in understanding the heterogeneous observed benefits without resorting to *post hoc* analysis. Determining the statistical significance of the difference in treatment effect between male and female patients is another example of such treatment-by-subgroup interaction effect analysis.

If analysis populations for primary effectiveness and safety endpoints are prospectively defined in a SAP, potential confusion in interpreting the study result can be avoided and hence a higher level of evidence established. To adequately characterize the analysis populations, such as intention-to-treat (ITT) or modified-ITT, in a SAP, a sponsor may include precise definitions of each analysis population and the statistical inferential methods planned for each.

To adequately describe the primary and secondary endpoint analyses, a SAP not only describes the statistical methods used for these analyses in sufficient detail, it also includes a plan to assess the validity of the assumptions on which these statistical methods are based. Examples of such assumptions include distributional assumptions for parametric inference and goodness-of-fit for model-based inference. If simulation methods, such as Markov Chain Monte Carlo, are conducted for statistical inference, model-checking diagnostics can confirm the validity of the simulation results.

If the primary and secondary endpoint analyses involve multivariate modeling such as regression, a SAP can provide transparency of the model-building process by prospectively identifying a list the candidate covariates (including time-dependent covariates) and the model building strategy (e.g., variable selection procedures and transformation).

In the case of a submission with an intention to make labeling claims based on secondary analyses, an appropriate plan in SAP to handle multiplicity for testing multiple hypotheses on secondary endpoints and subgroups will keep the inflation of overall Type I error rate under control and thus maintain the level of the evidence concluded from the study results.

Despite great effort undertaken at the study planning and conduct phase to collect all the data as specified in a protocol, there are almost always some missing values in study results. Since missing data can introduce potential bias, a missing data handling plan is an essential part of a SAP to establish the validity of study conclusions. An assessment of the reasons that cause the missing data, such as protocol violation or patient withdrawal due to adverse experience, provides important insights to the nature of missing mechanism and thus its potential impact to the trial. If the pattern of missing value appears to be random, some analytical techniques may be used to address issues of missing data. Although these techniques often employ major assumptions that cannot be fully validated for a particular study, the study protocol should pre-specify appropriate statistical data analysis methods for handling missing data. In addition, it is essential to investigate the sensitivity of the analysis results to various method of imputing missing values for all inferential endpoints. If a SAP contains a detailed list of covariates in the multiple imputation model and a justification of the model, a clear procedure will be in place at the end of the study when the data is analyzed.

The Statistical Analysis Plan (SAP) and the IDE

It is expected that an IDE contain a justification of the size of the study, the so-called sample size calculation. In order for that to occur it is crucial that the IDE identify the primary endpoint(s) for both safety and effectiveness. In the case of a study that uses hypothesis testing (most studies for therapeutic and aesthetic devices do so), a test statistic for providing the primary analysis should also be identified, since this would be used to make the sample size calculation. However, it may not be necessary to provide a detailed and complete Statistical Analysis Plan at the time of the IDE submission.

The natural question that then arises is when does the complete Statistical Analysis Plan needed to be finalized for a scientifically valid investigation. The answer depends on whether the investigation is blinded (masked) or not. If the study is double-blinded (double-masked) so that neither the patient nor the clinical study team knows the treatment assignment, then it is scientifically valid for the sponsor can submit to FDA modifications in the SAP up until the time of the data lock, provided no one has access to the treatment assignment codes. If the study is not blinded to the clinical study team, regardless of whether the patient is blinded or not, then the SAP needs to be locked down completely before the first subject reaches any primary endpoint in order to preserve the study's scientific integrity. If a safety or effectiveness endpoint can occur shortly after enrollment this implies that the SAP needs to be completely specified before then. In the time period between the approval of the IDE and this point the complete SAP needs to be submitted to the FDA as an IDE supplement.

What would need to be submitted to provide assurance that the study design has the statistical components to assure that it can if it turns out well result in a marketing application? This checklist can provide a guide to answering that question. It is important to note that up until the time that the blind is broken for double-blind studies or until any subject reaches an outcome for others, that a change in the SAP can be submitted as a supplement to the approved IDE. Changes to the study design and the SAP plan after that point severely threaten the scientific validity of the study.

Although a study's SAP is often part of the study protocol, sometimes the sponsor may choose to make SAP as a separate document from the study protocol as they may be finalized at different time. Unlike the study protocol of pivotal study, the SAP may not need to be finalized at the time of IDE submission if the pivotal study is a double-blinded. In other words, if neither patients nor the clinical study team knows the treatment assignment, then the SAP can be changed up until the time of the data lock, provided no one has access to the treatment assignment codes. If the study is not blinded to the clinical study team, regardless of whether the patient is blinded or not, then the SAP needs to be locked down completely before the first subject reaches any primary endpoint. If a safety or effectiveness endpoint can occur shortly after enrollment this implies that the SAP needs to be completely specified before then. In the time period between the approval of the IDE and this point the complete SAP needs to be submitted to the FDA as an IDE amendment supplement.

IDE Modifications

FDA believes that the following types of protocol changes would require an approved IDE supplement and SAP should be revised accordingly because they are likely to have a significant effect of the scientific soundness of the trial design and /or the validity of the data resulting from the trial:

- Change in indication,
- Change in type or nature of study control,
- Change in primary endpoint,
- Change in method of statistical evaluation, and
- Early termination of the study (except for reasons related to patient safety).

Supplements for Adaptive and Bayesian Submissions

In the FDA Critical Path Opportunities List released in March of 2006, under the Heading of Streamlining Clinical Trials, topic #36 is “The Use of Prior Experience or Accumulated Information in Trial Design”. This includes both Adaptive Trial Design and Bayesian (and other non-frequentist) Methods.
http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077258.pdf

The next two sections address considerations for these types of investigations at the IDE stage.

Supplemental Considerations for an Adaptive Design in an IDE Submission

In planning a pivotal study, it is sometimes extremely advantageous to consider an adaptive design. Often there is little data available to provide an adequate estimate of the variance, so that the estimated sample size under the fixed trial design can be quite off-target. Planned interim data analyses in an adaptive design can provide additional chances to re-size the trial appropriately. In such cases, the sponsor is encouraged to work with FDA statistician early and closely when contemplating an adaptive study design.

The proposed adaptive design study protocol should provide adequate details regarding “when”, “what” and “how” in adaptation. The “when” refers to when the adaptation is scheduled to take place. In the case of interim analysis this adaptation could be based on the information time, for example, at 50% information time, at which time an interim analysis will be performed. The “what” refers to the specific trial aspect that would be adapted. The “how” concerns detailed data analysis plan at interim look and possible actions for the later stage of the trial. A flow chart for decision rules among all possible decisions can be helpful in describing “how” adaptation is implemented.

The FDA statistical reviewer will examine the IDE submission to make sure that it provides adequate details to address some statistical concerns associated with utilizing adaptive trial designs. Those concerns include the possible inflation of Type I error rate due to increase of overall sample size, and bias associated with the multiplicity of testing options. Different adaptation strategies require different statistical considerations in final data analysis to address the bias issue and/or the increased false positive rate. Sample size re-estimation is one of the most popular kinds of adaptation in FDA submissions. If sample size recalculation is based on observed interim treatment effect, Type I error rate of the trial could be inflated three-fold under the revised sample size, and therefore statistical adjustment would need to be proposed and implemented in order to control the Type I error rate under the required level. For example, if the sample size is increased based on interim analysis, then a planned approach that down-weighs later data can be applied in final data analysis to provide tight control the Type I error rate. For some types of classical adaptive designs, such as the group sequential designs, statistical concerns on possible bias and Type I error rate have been well investigated in the literature and appropriate adjustments are available for implementation. It is expected that the protocol provide adequate citations of the literature on the planned adaptation methodology for such adaptive designs. For novel adaptive designs or for complex adaptive strategies, reference to published research alone in the protocol may not be sufficient in providing adequate support for a reviewer’s statistical concerns. In such cases, simulations of proposed adaptive trial design may also be necessary to justify the validity of the trial design. Simulations are essential for most Bayesian adaptive designs.

Trial simulations rely on assumptions of statistical models and the FDA statistical reviewer will question the reasonableness of those assumptions for the proposed study. It is expected that simulations will be presented in the IDE submission under a wide range of assumptions to cover different clinical trial scenarios. The operating characteristics of the trial assessed in simulations should include, but not limited to, the Type I error rate for the overall study (for safety or effectiveness), power of the trial design and sample size distribution if sample size is not fixed. An in-depth investigation regarding the impact on the simulated outcomes due to different random seeds, the number of replicates per simulation and different clinical trial setting should be conducted. The reviewer will check to see if computer programming codes of the simulation are included in the submission.

Interim looks at the data could introduce operational bias, especially when unblinded (unmasked) data are assessed. Operational bias is much harder to quantify than statistical bias, therefore it’s very important to include in the protocol some strategies to mitigate the possible operational bias. The protocol should detail who, among sponsor, independent statistical consultants, and data monitoring committee, have access to the unblinded data. The knowledge of interim data has potential to affect the ways of how patients are enrolled, treated, managed or evaluated, and therefore it is a great source of operational bias. A firewall may be needed to shield investigator/sponsor as much as possible from the knowledge of the interim data. (see the FDA guidance on Data Monitoring Committees for a discussion of this (FDA, 2006).)

If a Bayesian adaptive design is proposed, then many of the same items specified in the general discussion on adaptive design also apply for a Bayesian design. An extensive resource on Bayesian adaptive trial designs is the recent book by Berry et al (2010). Most important in an IDE with an adaptive design is a thorough description of the adaptive design, including what aspects of the design are potentially adapted (e.g., randomization ratio, sample size, etc.), number and timing of interim looks, decision rules at each look, stopping thresholds (for Bayesian designs, thresholds are often based on the predictive probability of study success), minimum and maximum sample sizes (if sample size is adapted), and the unit of information. Predictive probability is the probability of achieving success on the primary endpoint(s) when all subjects have completed follow-up. The probability includes all enrolled subjects, and can also include subjects yet to be enrolled, up to the maximum intended sample size. If predictive probability is used to predict outcomes for subjects yet to be observed, then the prediction model should be pre-specified in the IDE protocol. This is particularly important when the primary endpoint is binary, but follow-up time is available on enrolled subjects. A time-to-event model might then be used for prediction.

Simulation of the proposed adaptive trial design may be an appropriate method to justify the validity of the adaptive trial. For Bayesian adaptive designs, FDA is interested in understanding the Type I error rates and power for adaptive clinical trials under realistic scenarios. For a few frequentist and for many Bayesian adaptive trial designs, assessment of frequentist operating characteristics can be difficult to derive analytically. An in-depth investigation regarding the impact on the simulated outcomes due to different random seeds, the number of replicates per simulation and different clinical trial setting should be conducted. For example, an FDA statistical reviewer might request that sponsors submit simulations of operating characteristics under various assumptions, including the null hypothesis and main alternative hypothesis. Some additional assumptions include simulations under various amounts of borrowing from prior studies, different control means or rates (or variances), different treatment effects (including the null and alternative), different accrual rates, different correlations among primary endpoints (if applicable), different transition models among time points (if applicable), different randomization ratios, and different rates for loss to follow-up. A statistical reviewer might also request operating characteristics for an assumed treatment effect that is close to the null space, but technically within the alternative space, to ensure that a non-clinically meaningful effect is unlikely. Finally, computer programming code of the simulation should be included in the IDE submission.

Supplemental Bayesian Considerations in an IDE Submission

FDA’s Center for Device and Radiological health has led an effort to encourage the use of Bayesian statistics in medical device clinical trials. (See Campbell (2011) and Bonangelino et al (2011) for some history and examples.) It is important to note that since almost all Bayesian studies are adaptive, the considerations in the above section on adaptive studies and in the accompanying Adaptive Supplement can and do apply to most

Bayesian studies. The Bayesian checklist supplement addresses additional considerations (besides the adaptive ones) for a Bayesian submission.

For a Bayesian study, there are different recommendations depending on whether the study incorporates objective prior information (from previous studies, perhaps) or not. If a study incorporates objective prior information, then the statistical reviewer will look for a description of the prior studies that a sponsor proposes to incorporate, including the level of data available from each prior study (e.g., patient-level data, study-level data). A reviewer will also look for a qualitative assessment of exchangeability of the prior studies with the current study, with respect to the endpoints of interest. This will likely involve a clinical argument, in addition to perhaps a statistical argument. Statistical reviewers also look for a description of the model used to incorporate the prior data, as well as an assessment of the prior influence or borrowing strength. For the latter, the prior probability of the claim might be provided, an evaluation of the Type I error rate might be provided, or the prior effective sample size could be determined. The prior effective sample size is the effective number of patients expected to be borrowed from prior studies. The statistical reviewer and sponsor statistician should discuss what information is needed to assess prior influence. Finally, in order to borrow strength from previous studies, covariate information might be needed to calibrate the prior studies with the proposed study. The covariates should be included within the model used to borrow strength from the prior studies. Pennello and Thompson (2008) provide a detailed discussion of Bayesian submissions for medical device trials.

Subjective priors are typically not recommended for parameters associated with the primary endpoint. However, for hyper-parameters, subjective priors are often used. Statistical reviewers will ask for justification of any subjective priors, and check for the extent of their sensitivity and influence on study claims. Subjective priors are also routinely used as design priors for adaptive designs. For example, a subjective prior might be used to estimate sample size for a trial (i.e., stopping for accrual). Description of these priors will help the statistical review of an IDE submission. Irony and Pennello (2001) provide additional details of priors for medical device studies.

For Bayesian submissions, FDA is interested in controlling Type I error rates and calculating statistical power for medical device clinical studies. Because of the incorporation of prior information, the Type I error rate of many Bayesian designs can be higher than the customary rate for frequentist designs, which do not formally incorporate prior information. Statistical reviewers therefore will expect to see an assessment of the Type I error rate for Bayesian designs in order to evaluate whether the increase in the error rate is appropriate clinically. Assessment of frequentist operating characteristics is discussed in more detail in the section on adaptive design.

Finally, various modeling issues are unique to Bayesian analyses because of the routine use of MCMC sampling. Thus, the FDA statistical reviewer will look to see in the IDE that appropriate model checking and convergence methods be planned, especially if the likelihood model is complicated. In addition, a description of the imputation of missing data is helpful. The FDA's "Guidance for the Use of Bayesian Statistics in Medical Device Trials" (FDA, 2010) describes in more detail the information a statistical reviewer might recommend in an IDE for a Bayesian study.

References

Berry, S. M., Carlin, B.P., Lee, J.J. & Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials.* CRC Press, Boca Raton, FL.

Bonangelino, P., Irony, T., Liang, S., Li, X., Mukhi, V., Ruan, S., Xu, Y., Yang, Y. & Wang, C.G. (2011). Bayesian approaches in medical device clinical trials: A discussion with examples in the regulatory setting. *Journal of Biopharmaceutical Statistics* **21**: 938-953.

Campbell G. (2011). Bayesian statistics in medical devices: Innovation sparked by FDA. *Journal of Biopharmaceutical Statistics.* **21(**5):871-887.

Food and Drug Administration (2006). The Establishment and Operation of Clinical Trial Data Monitoring Committees for Clinical Trial Sponsors. Guidance for Clinical Trial Sponsors - Establishment and Operation of Clinical Trial Data Monitoring Committees (finalized 3/27/06) http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf (accessed Sept. 27, 2012).

Food and Drug Administration (2010). Guidance for the Use of Bayesian Statistics in Medical Device Trials (finalized February 5, 2010. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm (accessed Sept. 27, 2012).

Food and Drug Administration (2011). Draft Guidance for Design Considerations for Pivotal Clinical Investigations for Medical Devices (issued August 15, 2011). http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM267831.pdf (accessed September 27, 2012).

Irony, T.Z. & Pennello, G.A. (2001). Choosing an appropriate prior for Bayesian medical device trials in the regulatory setting. In *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*. Alexandria, VA: American Statistical Association.

Levenson, M.S. & Yue L.Q. (2013). Regulatory Issues of Propensity Score Methodology Application to Drug and Device Safety Studies. *Journal of Biopharmaceutical Statistics* **23**(1) (to appear).

Pennello, G.A. & Thompson, L.A. (2008) Experience with Reviewing Bayesian Medical Device Trials, *Journal of Biopharmaceutical Statistics* **18**(1):81-115.

Vishnuvajjala, R.L. (2007). Statistical Issues in Diagnostic Devices Including

ROC Methods. In *American Statistical Association 2006 Proceedings of the Statistics in Epdiemiology Section*. Alexandria, VA: American Statistical Association.

Vishnuvajjala, R.L. et al (2012). Investigational Device Exemption Quality Review Checklists for Diagnostic Submissions at FDA/CDRH. In *American Statistical Association 2012 Proceedings of the Biopharmaceutical Section*. Alexandria, VA: American Statistical Association.

Yue, L.Q. (2007). Statistical Review Quality Assessment for Therapeutic PMA Submissions. In *American Statistical Association 2006 Proceedings of the Biopharmaceutical Section*. Alexandria, VA: American Statistical Association.

2006.
Yue, L.Q. (2007). Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *Journal of Biopharmaceutical Statistics* **17**: 1-13.

Yue, L.Q. (2012). Regulatory Considerations in the Design of Comparative Observational Studies Using Propensity Scores. *Journal of Biopharmaceutical Statistics* **22**(6) (to appear).

IDE Statistical Quality Review Assessment (“Checklist”) for Therapeutic and Aesthetic Submissions

I. Introduction/Background

1) Is the purpose (primary objective) of the study clearly stated?

2) Is the Intended Use/Indications for Use clearly stated?

3) Are results from previous feasibility/pivotal studies and on-going studies (US/OUS) clearly described?

4) Are the regulatory history, pre-IDE submission and meetings reported in the submission?

II. Study Design

1) Is the type of study design (e.g., randomized, non-randomized with concurrent or historical control arm, single arm study) clearly identified?

2) Is each intervention (e.g., the therapy each patient receives) clearly described?

3) Is there a detailed description of treatment assignment and randomization scheme (e.g., randomization ratio, fixed or adaptive; blocking or stratification variables)?

4) If non-randomized study has a control, is the source of the controI group clearly explained along with how the bias will be minimized and group comparability addressed?

5) Have the screening criteria (and its timing relative to randomization), enrollment schedule, study duration, and follow-up schedules (including time and places for follow-up, primary endpoint measures to be taken at each follow-up visit) been clearly specified?

6) Is there a plan to prevent missing data, by minimizing loss-to-follow-up and other protocol deviations?

7) Are the study centers clearly identified? (maximum number of centers, US/OUS, minimum and maximum number of subjects in each center)

8) If blinding is appropriate, is there a scheme to blind individual patient outcomes from the patient, evaluator, investigator, operator, and adjudicator? Is there a plan to blind aggregated outcome data during the study? Is there a plan to assess blinding?

9) Is there a plan for monitoring the quality of the data collection?

III. Primary and Secondary Endpoints

1) Are the measurement and evaluation of the primary and secondary endpoints (for effectiveness and safety) clearly defined at the patient level?

2) Are any performance goals clearly stated and justified? (e.g. consideration of comparability issues: relevance of PG to current study)

3) Are there mathematical expressions and verbal statements of the hypotheses to be tested?

4) Is the significance level specified as 1-sided or 2-sided for each hypothesis, or if Bayesian is threshold for posterior/predictive probability specified?

5) Are the non-inferiority margins and/or superiority margins, if any, specified and justified?

6) If both non-inferiority and superiority are contemplated, is there a plan for simultaneous testing of superiority and non-inferiority?

7) For multiple testing (if any), is there specification of statistical methods to handle multiplicity for and overall control of Type I error rate?

8) Have the criteria for overall study success been clearly specified? (hypotheses or other observed result involved in the determination of study success)

9) Is a minimum sample size for safety endpoint(s) specified, if appropriate?

IV. Sample Size Determination

1) Is the sample size (and power) correctly calculated? And are Type I and Type II error rates controlled?

2) Are the expected effect size(s) and assumptions under which the sample size is calculated justified? (e.g., control rates, variability assumption, correlation among endpoints, and other model assumptions)

3) Have possible missing data been taken into consideration in the sample size calculation?

4) Could there be statistical significance without clinical significance with the sample size?

V. Statistical Analysis Plan

1) Are the baseline covariates including demographic information and time-dependent covariates to be measured on subjects clearly identified?

2) Are important subgroups identified and their planned analyses described?

3) Concerning baseline covariate assessment, is there a plan to assess balance between interventions, data poolability across sites, geographic regions and important subgroups? If so, is there an alternative analysis plan if data are not poolable across sites/regions/subgroups?

4) Concerning the primary analysis populations for primary effectiveness and safety endpoints

a. Are the appropriate analysis populations clearly defined? (e.g., intention-to-treat (ITT), modified ITT, As-treated, Per-Protocol, etc.)

b. Has the statistical methodology for hypothesis testing within each appropriate analysis population been clearly specified?

c. Concerning the assessment of model assumptions, are the distributional assumptions reasonable and are the parameters identifiable?

   d. Are model checking diagnostics supplied if simulation is done (e.g., MCMC)?
   e. Have multiplicity issues for testing multiple hypotheses for secondary endpoints been adequately addressed?

5) Missing data handling
   a. Is there a plan to assess missing data due to dropout or other protocol deviation?
   b. Is the handling of missing data in covariates and in clinical outcomes adequately addressed?
   c. Have sensitivity analyses of the missing data for all inferential endpoints been planned?
   d. If there is an imputation model, are the covariates for that model identified and are the assumptions justified?

6) Is there a plan for data lock and electronic submission of patient-level data? Links to preferred format of submission of electronic data
   a. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/ucm134508.htm
   b. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/ucm136377.htm
   c. Has a statistical analysis strategy to handle treatment variations been pre-specified in the protocol?

7) Concerning the primary and secondary endpoint analysis for safety and effectiveness, are analyses for important subgroups (e.g., gender) identified (in particular, significance levels for subgroup-by-treatment interactions)?

Supplement for Adaptive Designs

1) Are the design features for the adaptive design clearly specified? These would include:
   a. Aspect of adaptation: early stop for effectiveness, early stop for futility, sample size re-estimation (early stop for accrual), adaptive selection of subgroup, randomization ratio (e.g., play-the-winner), drop-the-losing-arm based on interim results, population enrichment
   b. Is the method by which the adaptation takes place adequately described?
   c. Blinding issues
      i. Will unmasked information be used at the interim looks (e.g., treatment effect)? Or will only blinded information be used?
      ii. Which entities remain blinded at interim looks?
   d. If the interim analysis is based on an intermediate or surrogate endpoint (e.g., an outcome with a shorter follow-up time, does the submission provide evidence that this is reasonable?
   e. For interim analyses, is the timing and number of the looks based on information unit and timing/number of the interim looks and are the test statistics and decision rules clearly specified at each look and at the final analysis?

2) Statistical considerations
   a. Is the study-wide Type I error rate well-controlled?
   b. In the case of sample size re-estimation, are the maximum, minimum and average sample sizes reported?
   c. Has the statistical bias in estimates of treatment effect and in confidence intervals been addressed?

- **d.** Has the potential for increased Type II error rate or decreased power for each study hypothesis been considered?
- **e.** Have details of analytic derivations been provided, if appropriate?
- **f.** Has published literature been provided for support of these statistical considerations?
- **g.** Have simulations been provided to estimate the operating characteristics of the design (if necessary), including the Type I error rate under a range of parameter values for the null hypothesis, the power under a range of clinically possible parameter values for the alternative hypothesis, and the sample size distribution?
- **h.** Has there been a comparison of the adaptive designs to the non-adaptive (fixed) design?
- **i.** Have the computer programs for the simulations been submitted?

3) Logistical considerations

- **a.** If there is a Data Monitoring Committee (DMC), is there a written charter of DMC, including the reporting structure to the sponsor and Steering Committee?
- **b.** Have operating procedures, firewalls, and written agreements regarding who performs the unblinded analyses and sponsor/CRO involvement in recommendations for adaptations been clarified?
- **c.** Have entities been specified who would remain blinded from the result and decision of interim looks?
- **d.** Has the communication between the sponsor, FDA and DMC regarding the interim results (meeting minutes) been specified?
- **e.** Does the submission indicate how are data obtained to conduct interim analyses?
- **f.** Will subject enrollment be temporarily on hold during the interim analysis awaiting DMC recommendations?

Supplement for Bayesian studies

Since almost all Bayesian studies are adaptive, the considerations in the above section on adaptive studies would apply. This supplement addresses additional considerations for a Bayesian submission.

1) Incorporation of Objective Prior Information

- **a.** Is the description of prior studies complete (including primary endpoints, protocols, treatment groups, patient populations)?
- **b.** Has the level of data available from each study (e.g., patient-level data , summary statistics by study groups) been clearly specified?
- **c.** Is a clinical assessment made of exchangeability of treatment effects across prior and current studies?
- **d.** For informative priors, has the prior influence or borrowing strength been assessed?
    - i. Is there clinical and statistical justification of the strength to be borrowed via the model (e.g., if pooling, justification of the discount rate, if using a hierarchical model, justification of the hyper variance) and a justification of the threshold for the Type I error rate?
- **e.** Has a description of the plan of how to calibrate prior studies with proposed study, using measured covariates if appropriate, been submitted?
- **f.** Has the model proposed to incorporate prior data (e.g., hierarchical, power prior, commensurate prior) been adequately described?

2) Regarding Operating Characteristics (for each endpoint), did the sponsor include simulations of Type I error rate and power:

a. Assuming various amounts of borrowing from prior studies, if applicable?
b. Assuming different control means or rates, or different control variances?
c. Assuming different treatment effects?
d. Assuming different accrual rates?
e. Assuming different randomization ratios, if applicable?
f. Assuming different correlations among primary endpoints, including independence? (if applicable)
g. Assuming different transition model or correlation among time points?
h. Assuming different rates for loss-to-follow-up?

3) Modeling Issues

a. Will Monte Carlo simulation be needed (e.g., MCMC)? Does the sponsor plan model checking or convergence methods?
b. Are parameters of interest identifiable?
c. Are the non-informative priors or subjective hyper-priors clearly described?
d. How sensitive are the study results to the choice of hyper-priors?
e. For missing data, have an imputation model and a sensitivity analysis been proposed in the submission?