Quality Statistical Review Checklist of Investigational Device Exemption (IDE) Submissions for Diagnostic Medical Devices

R. Lakshmi Vishnuvajjala, Shanti Gomatam, Gregory Campbell, Yonghong Gao, Gene Pennello, Kyunghee Song, Laura Thompson, Lilly Yue, 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 Food and Drug Administration (FDA) that, if approved, allows a sponsor (i.e., an investigator or a device company or some other entity) to begin an investigational study on humans in the United States for any study that poses potential significant risk to the subject. The time limit for FDA to take an action such as approval, approval with conditions or disapproval is 30 days. If a study does not pose significant risk then no IDE would be required.

As part of its internal quality review procedures, and similar to the effort for Pre-Market Applications (PMAs) in the past (Yue, 2007; Vishnuvajjala, 2007), the Division of Biostatistics (DBS) has developed "IDE checklists" to aid its reviewers in carrying out complete and consistent reviews of IDEs. These reviews are carried out from the perspective of assessing whether the clinical investigations and supporting studies could potentially support a premarket submission. A version of this checklist for diagnostic devices is attached in the appendix of this paper. The therapeutic version of this checklist can be found in the companion paper by Campbell et al (2012). While these checklists contain elements considered to be critical for an IDE review, they may not be comprehensive, nor is every point raised here applicable to every device.

Such checklists can also be beneficial to statistical counterparts in the device industry by providing transparency regarding FDA's considerations in its statistical review of an IDE. The Draft Pivotal Clinical Investigation guidance (FDA, 2011b) can also be a useful resource for sponsors designing studies for devices. Sponsors who have questions on exclusions or other specifics are encouraged to contact CDRH via the pre-submission process.

'Diagnostic' in this checklist is used to differentiate from 'therapeutic' or 'aesthetic'(see Draft Pivotal Clinical Investigation Guidance (FDA, 2011b) for further explanation) and is intended to be applied to all devices or tests whether they are used for screening, diagnosis, monitoring etc. However, this diagnostic checklist is not intended to cover "companion diagnostics" (i.e., diagnostic devices that a corresponding therapeutic product depends on for its safe and effective use; see FDA, 2011a).

IDEs are relatively rarer for diagnostic devices compared to therapeutic devices, especially for *in vitro* diagnostics, as diagnostic devices are less likely to be significant risk devices. Review concerns highlighted here for IDEs are also pertinent for diagnostic submissions that do not require an IDE. Statistical reviewers are likely to use elements of this checklist to guide their reviews even when reviewing such protocols.

The FDA Safety and Innovation Act (FDASIA) has incorporated some important changes to the Food Drug and Cosmetic Act. Section 520(g)(4)(C) now states 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. It 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.

The following sections address elements of the Diagnostic Checklist.

I Background

A statistical reviewer of diagnostic devices would usually begin by considering the background of the target condition (disease or condition of interest being assessed by the device), and the device description including: technology or principle of operation, the nature of the clinical claim (e.g., screening, diagnostic, measurement), thresholds/cutoffs used to make clinical determinations about the target condition, device inputs and reports (including reports of equivocal results).

Understanding the intended use of the device is a critical part of this background review. For a study involving a diagnostic device, the target condition (disease or condition of interest), the intended subjects, the intended users and intended setting need to be clearly identified. The same device can be used in a variety of settings by different sets of users – differences in intended user/setting and subjects may results in different device requirements and level of scrutiny, and could thus lead to different concerns in the statistical review. For example, when used for screening a large population of asymptomatic subjects, a screening test is often expected to have high specificity to avoid an inordinate number of false positives. Similarly, users of the device could be professional laboratory technicians, staff in a medical facility, technicians in a CLIA certified laboratory, or the consumer who purchases an over-the-counter kit – ease of use and accurate reporting would be considered more critical for the over-the-counter use by an untrained layperson for instance.

Location or source of the specimen (blood, interstitial fluid, urine, specimen from tumor tissues and saliva or some tissue samples), and other details like specimen type, matrix, medium, size, fixative etc. are reviewed because these factors can make a difference to device performance and use claims.

Regulatory history and background on previous submissions is also considered in a statistical review; however, the intent of this checklist is to focus on statistical design and analysis issues.

II Pre-clinical and Other Studies

Clinical development of a device can encompass pre-clinical, animal, feasibility and other exploratory studies. Typically a diagnostic IDE proposes a collection of supporting

studies in addition to the pivotal clinical investigation. While all such studies may not necessarily be reviewed in detail by a statistical reviewer, the existence of such studies and evidence provided by them is noted in the process of doing a review. On the other hand, for several diagnostic devices statistical review of certain pre-clinical studies, or proposals for such studies, is an important component of the statistical review. For example, statistical review of analytical specificity, precision studies (repeatability/reproducibility studies) that evaluate device precision under critical sources of variation, and limit of detection, are usually carried out. Depending on the device, various bench and laboratory tests may also need to be reviewed. When diagnostic devices are applied to specimens, stability issues might have to be statistically reviewed.

Of particular concern to the diagnostic statistical reviewer are studies that may be used for cut-off or classifier determination. It is useful for any such "training" studies or procedures to be explained in detail in the protocol – it is critical that the validation of a device be carried out on data independent of that used for its training.

III Pivotal Clinical Study Design

Review of the pivotal clinical study design encompasses many aspects. The purpose of the study and its primary objectives are critical to the review; these are expected to be stated, and to be consistent with the intended use of the device. A reviewer would look for clear statements of primary clinical endpoints or primary diagnostic performance measures with associated hypotheses or estimation goals, and of the study success criteria. Adequacy of clinical and other justification for such goals would be evaluated collaboratively within the review team.

Safety and effectiveness for a diagnostic device are often not separable; diagnostic performance measures or primary objectives may contain elements of both safety and effectiveness. For the same reason, no single objective or measure may capture all aspects of either safety or effectiveness. A particular concern for a diagnostic reviewer is whether the collection of primary objectives and study success criteria adequately capture the "diagnostic trade-off". For instance, goals regarding false-positives *and* false negatives are both critical components of the study success criteria for a device intended for diagnosis. The sensitivity of a device can be improved at the cost of its specificity; both aspects of its performance are relevant when making a decision on device performance. Similarly bias and precision of a measuring device would both be of concern.

Descriptions of study design (paired, parallel groups, cross-over etc.), listing of number, location and type of study sites, and selection method for study subjects, screening and enrollment criteria for study subjects, storage and handling of subject specimens when applicable, are all reviewed with a view to assessing whether the study population would represent the intended use population and/or introduce sources of bias for the assessment of device performance. Comparator devices or methods are checked for clinical acceptability with the review team, as is clinical reasonableness of the washout period for paired or cross-over designs.

In diagnostic performance studies the sponsor's choice and pre-specification of clinical reference standard (CRS; refer FDA, 2011b; also referred to as "reference standard" in FDA, 2007) is critical and has to be clinically acceptable. Sometimes several defining details are needed for the CRS – for instance, when using an expert panel to define a CRS,

criteria for choice of panel members and their qualifications, clear guidelines on how calls will be made, and, if applicable, procedures to deal with cases where there may not be clear-cut decisions, are all required to be pre-specified with clarity and detail.

Diagnostic devices often involve readings and/or interpretations by personnel who may be referred to as operators, readers, evaluators, or by other medical designations (for example, pathologists for optical microscope readings). The qualifications and training provided to these personnel to train them in device use are vital elements of the study design. Because of the need for such personnel for some diagnostic devices masking/blinding in diagnostic studies can present interesting challenges and opportunities. Masking can be challenging because it has to consider evaluators of the CRS, the investigational device, and any comparators, in addition to the principal investigators/treating physicians and the subjects. It presents additional opportunity over that for therapeutic or aesthetic devices because it may be possible to blind/mask the evaluators of the different modalities even when it is not possible to mask subject and treating physician to device use. Blinding of evaluators to results across modalities is an important component of the study design. Details on all of these aspects, listing any procedures to be used, are critical to the statistical reviewer. Any plan that assesses whether masking was successfully carried out provides additional useful information.

The statistical reviewer reviews the protocol for details of study duration, any time-related issues (dependence or separation of evaluations by different modalities for instance) protocol deviations, and adverse event identification and reporting.

IV Basic Statistical Analysis Plan

The basic statistical analysis plan is reviewed before the reviewer verifies sample size calculations, since elements in the basic analysis plan are critical components of the design that are required for sample size calculation.

Statements of primary and secondary clinical endpoints or diagnostic performance measures are the core elements of a Statistical Analysis Plan. Statistical reviewers look for explicit definitions of endpoints and all terms and symbols used. Clear statements, preferably written out both mathematically and verbally, help the reviewer understand hypotheses or estimation goals. A reviewer would check whether the study success criteria based on these endpoints or diagnostic performance measures with associated hypotheses or estimation goals support the intended use and facilitate evaluation of clinical consequences.

When studies have multiple hypotheses or estimation goals that need not all be satisfied for study success, a reviewer would check that appropriate adjustments for control of overall type I error probability (with acceptable type II error probability) are being applied.

The reviewer would need to obtain a clear understanding of: the statistical analysis methodology being used for estimation or testing of hypotheses; the assumptions that such analyses would be based on; whether the assumptions being made are justified and/or will be assessed; whether there is an alternative plan if assumptions do not hold; whether analysis populations (e.g., intent-to-diagnose, per protocol) have been clearly specified for each of the proposed analyses and are clinically reasonable. The reviewer

would ascertain that critical covariates are included in models, and that important subgroup analyses have been specified.

Potential biases such as selection and spectrum bias, verification bias (when not all subjects have a CRS measurement), the bias due to an imperfect CRS, and sponsor's plans to avoid or handle such biases would be considered in the review. The sponsor's plan for handling missing data and invalid data in the endpoints, diagnostic performance measures, or covariates would be evaluated with associated assumptions and sensitivity analyses.

V Sample Size Determination

When assessing study sample size a diagnostic reviewer evaluates whether the proposed sample size is consistent with the stated study success criteria and statistical analysis plan. As stated earlier, the study success criteria for diagnostic devices are checked to see that they cover the "diagnostic trade-off". For devices providing measurements, reviewers evaluate whether the study is adequately powered for any agreement assessments, taking into account potential measurement error. Hypothesis testing or estimation goals for the primary endpoints or diagnostic performance criteria drive such sample size calculations, with statements of type I and type II error levels intended to be achieved, statements of, and adequate and clinically acceptable justification provided for, any estimates used to make these calculations. Allowances made for missing data are evaluated, with attention to adequate upward adjustments of sample size as necessary.

For fixed sample size designs, the reviewer verifies the sample size that covers all requirements (typically the maximum over calculations for all required primary objectives) for adequate study sizing. For studies with sequential designs and Bayesian or frequentist adaptive designs sample size is not fixed upfront but depends on interim results observed. Sample size for sequential designs would be reviewed for adequacy against appropriate methodology. Additional review considerations for Adaptive and Bayesian designs are discussed in subsequent sections.

VI Detailed Statistical Analysis Plan

While evaluating the detailed data analysis plan a reviewer would check whether the Statistical Analysis Plan incorporates reporting of subject/specimen accountability, demographic information, baseline characteristics or any related clinical characteristics. Reviewers would also check what the sponsor plans to do if there are imbalances in important characteristics.

Sponsor's plans for assessing poolability across sites or US/OUS sites are reviewed with the understanding that for some diagnostic clinical investigations study sites that are expected to differ in study subject composition are specifically chosen to provide adequate representation of the spectrum of disease and to sufficiently cover the intended use population. (For instance, collection and testing sites usually provide different mixes of subjects. In such cases one does not expect poolability across sites.) When sites intended to be poolable are found to be not poolable, what does the sponsor plan to do?

A review of the detailed plan for addressing relevant biases like selection bias, verification bias etc. would be carried out.

Finally, it is important to assess that the Statistical Analysis Plan will be finalized either before study data become available or before the data are unblinded.

Supplement for Adaptive Designs

Medical product development has become increasingly challenging, inefficient, and costly. FDA's Critical Path Initiative (CPI) is the agency's national strategy to drive innovation into the scientific process. A main objective of CPI is to call attention to the need for more scientific efforts and efficient evaluation tools in medical product development. Adaptive study design is a topic in the medical utility area of CPI. Compared with traditional studies with fixed designs, studies designed with adaptive features may be more efficient (e.g., shorter expected duration, fewer expected subjects) and more likely to demonstrate medical product effectiveness and safety. The main motivation behind an adaptive design is that it allows a medical product developer to respond to what is learned in a study about their product and modify the study accordingly in mid-course, without compromising study integrity and validity.

An adaptive design demands more effort than a fixed design. Thus, in an IDE or a pre-submission, reviewers expect to see some clinical and/or statistical reasoning for utilization of an adaptive design. One reason might be that for a fixed design the sample size is likely to be far from correct due to uncertainty in the assumptions used to determine it. For example, at the design stage, the assumed variability of the study data may essentially be a guess because little or no prior data may be available to estimate it. In contrast, an adaptive design can provide a chance to re-size the study at an interim stage based on the interim estimate of data variability. When considering an adaptively designed study, a sponsor is strongly encouraged to schedule a pre-submission meeting with FDA and work closely with the FDA statistician.
A thorough description of a proposed adaptive design in an IDE or pre-submission would help FDA reviewers understand and assess the study design. The description would include all aspects of the design that are potentially adaptable and the rules for making the adaptive decisions, For example, in an adaptive sample size design, the IDE could describe the number and timing of the interim looks and the decision rules at each look (e.g., the stopping thresholds). A flow chart delineating the adaptive decision rules can be very helpful in describing "how" adaptations will be implemented.

FDA expects to see the statistical concerns of adaptive designs addressed in the study protocol. Concerns include possible type I error inflation and estimation bias due to the adaptive decision procedure. Statistical considerations depend on the adaptation strategy employed.

Sample size re-estimation is a common adaptive feature seen in CDRH submissions. A statistical concern is that an interim sample size recalculation based on an observed performance estimate or its variance could potentially inflate the type I error rate of a study. A statistical adjustment can be implemented to ensure that the type I error rate is controlled at the nominal level. One type of statistical adjustment for controlling the type I error rate relies on down-weighting later stage data in the final data analysis. While it achieves the objective of controlling type 1 error, down-weighting data is generally discouraged because it violates the likelihood principle.

For diagnostic device studies, sample size re-estimation and group-sequential designs are discussed in Tang and Liu (2010) and Tang, Emerson, and Zhou (2008). Besides sample

size, potential adaptive features of a diagnostic device study might include an adaptive diagnostic signature (classifier), an adaptive threshold in the device result for the purpose of making a binary diagnosis, and adaptive enriching a study with subpopulations in whom the diagnostic device is performing the best. Unfortunately the FDA has relatively little experience with these adaptations. Because the majority of diagnostic device studies are not randomized, some adaptations common in therapeutic device studies may not be applicable (randomization ratio, treatment arm dropping, etc.).

An adaptive study may also be at risk of operational bias due to some data having to be unblinded/unmasked at interim stages of the study to permit adaptive decision making. Operational bias is much harder to quantify than statistical bias. It is prudent to include in the protocol strategies for mitigating possible operational bias. The protocol of an adaptive study is expected to detail who, among sponsor, independent statistical consultants, and data monitoring committee (if applicable) will have access to the unblinded data. The knowledge of interim data has the potential to affect how subjects are enrolled, treated, managed or evaluated, and therefore can be an important source of operational bias. A firewall may be needed to shield investigator/sponsor as much as possible from the knowledge of interim data.

Many of the items in the above general discussion apply to Bayesian adaptive designs. Depending on what is being adapted, a Bayesian adaptive decision may be based on predictive probabilities. For example, suppose a diagnostic device result at baseline is claimed to be associated with presence or absence of a condition by a particular follow-up time. A rule for stopping early to declare study success could be based on a high predictive probability that the study would be successful if it was continued until all follow-up is completed on all subjects for the maximum sample size allowed. Such a predictive probability may consider the predictive distribution of missing condition status conditional on baseline covariates and early observations as well as the device result.

In a Bayesian adaptive study, the prior can be highly influential at an interim stage with a small sample size. Thus, before the first interim look is made a minimum sample size and minimum study duration may be enforced to ensure that sufficient data are collected to combine with the prior information. Notably, a *design prior* may be used when making a decision to stop enrollment, while a (usually less informative) *analysis prior* may be used for the final analysis after all data are collected. For further reference on Bayesian adaptive medical device studies, see "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials" (FDA, 2010) and Berry et al. (2010).
For adaptive designs that are well understood in the clinical study community, such as the group sequential design, statistical concerns such as type I error rate and estimation bias are well investigated and appropriate adjustments are available to implement. For well understood adaptive designs, the protocol can provide citations of the literature regarding the planned adaptation methodology.

For novel or complex adaptive designs, the reference of a published research paper alone is often insufficient for the reviewer to adequately assess statistical concerns, such as type I error rate. For example, the null distribution of the test statistic may not be known analytically. Frequentist operating characteristics may have to be simulated. For example, type 1 error rate simulation is commonplace for Bayesian adaptive designs due to their complexity.

Study simulations rely on assumptions of statistical models. Hopefully, the assumptions are clinically reasonable for the current study. Simulations may explore a wide range of assumptions to provide an in-depth investigation regarding the impact of different clinical study scenarios, parameter value assumptions, and adaptive decisions. An assessment of of a study can include its operating characteristics (type I error rate, power) and the distribution of sample size (if adaptive). In a Bayesian study, the distribution of the amount of borrowing from prior information (as measured by effective sample size) can also be obtained by simulation. A statistical reviewer usually prefers that the simulations cover a variety of scenarios, e.g., the null hypothesis, main alternative hypotheses, values of nuisance parameters (e.g., comparator device performance, data variance, correlation among endpoints, transition models among time points), and other study attributes (e.g., accrual rate, rate of loss to follow-up).

A statistical reviewer might also look at operating characteristics for parameter values close to the null space, but technically within the alternative space, to ensure that device performance that is non-clinically meaningful is unlikely to lead to a successful study. Simulations may be summarized in table format. For example, for each scenario used in the simulations, columns can display probability of study success, total study time, average sample size, probabilities of early study success and futility (at an interim look), and number of simulations of scenario. Reviewers could verify the simulation results if the IDE submission includes the computer programming code used for the simulations.

Supplement for Bayesian Considerations

For a Bayesian study, recommendations depend on whether the study incorporates prior information by objective or subjective means. An *objective* prior is defined here as one based on prior study data. All other priors can be considered *subjective* in the sense that they are formulated without the use of prior study data.

For sponsors proposing to incorporate objective prior information, statistical reviewers will look for a description of the prior studies, including the level of data available from each prior study (e.g., subject-level data, study-level data). The prior studies may also be checked for possible selection bias. Reviewers will also review the model used to incorporate the prior information. To borrow strength from prior studies, a model has to assume, to some extent, that the prior studies are *exchangeable* with the current study. Studies are exchangeable in a parameter if, prior to seeing the data, any ordering of study-specific parameters is considered equally plausible. Reviewers will look for a qualitative assessment of exchangeability of the prior studies with the current study, with respect to the parameters of interest. The assessment of exchangeability will likely involve a clinical or technological argument, in addition to perhaps a statistical argument. For diagnostic devices, the exchangeability assessment is at least bivariate in that a pair of performance parameters, e.g., sensitivity and specificity, is usually considered to investigate the trade-off of false negatives and false positives.

Statistical reviewers will also look for a description of the model used to incorporate the prior data, as well as an assessment of the prior influence or strength borrowed from the prior studies. Assessments that might be provided include the prior probability of the primary claim to be demonstrated, an evaluation of the type I error rate, and the prior effective sample size. The prior effective sample size is the effective number of subjects expected to be borrowed from prior studies. The statistical reviewer and sponsor statistician can discuss what information is needed to assess prior influence. To borrow

strength from prior studies, subject-level covariate information might be needed to calibrate the prior studies with the proposed study. The covariates can 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 diagnostic medical device studies.

Subjective priors, i.e., priors not based on data, are typically not recommended for parameters associated with primary safety or effectiveness endpoints. An exception is the use of a *non-informative* prior for a parameter, i.e., a diffuse prior that essentially gives no preference to any of its possible values. (In the statistical literature, some types of non-informative priors are referred to as objective priors.) Also, for *hyper-parameters* describing the distribution of a parameter of interest (e.g, the between-study variation in a parameter), subjective priors could be appropriate. 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 commonly used as design priors for adaptive designs. For example, a subjective prior might be used to estimate sample size for a study (i.e., stopping for accrual). Descriptions 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.

Although in theory, the Bayesian approach is not concerned with frequentist operating characteristics, the FDA is interested controlling the frequency with which it makes errors in regulatory decisions. Therefore, clinical studies, whether Bayesian or not, are generally designed to maintain a low false positive study rate and high power. Because of the incorporation of prior information, the false positive study rate of many Bayesian diagnostic designs can be higher than it would be if the prior information was not formally incorporated. Statistical reviewers therefore request an assessment of the error rates for Bayesian designs in order to evaluate whether the increase in error 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 Markov Chain Monte Carlo (MCMC) sampling. Thus, CDRH requests that appropriate model checking and convergence methods be planned, especially if the likelihood or prior model is complicated. In addition, a description of Bayesian imputation of any missing data is helpful. The FDA's "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials" (FDA, 2010) describes in more detail the information a statistical reviewer might recommend in an IDE for a Bayesian study. In addition, the text by Carlin and Louis (2010) provides a comprehensive overview of applied Bayesian methods.

References

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

Campbell G. et al (2012). Quality Statistical Review Checklist of Investigational Device Exemption (IDE) Submissions for Therapeutic and Aesthetic Medical Devices. In *American Statistical Association 2012 Proceedings of the Biopharmaceutical Section*. Alexandria, VA: American Statistical Association.

Carlin, B. and Louis, T. (2010). *Bayes Methods for Data Analysis*, 3rd Edition, CRC Press, Boca Raton, FL.

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 September 27, 2012).

Food and Drug Administration (2007). Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests (issued March 13, 2007). http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm (accessed September 27, 2012).

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

Food and Drug Administration (2011a). Draft Guidance for In Vitro Companion Diagnostic Devices (issued July 14, 2011). http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm262292.htm (accessed September 27, 2012).

Food and Drug Administration (2011b). 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.

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 Epidemiology Section*. Alexandria, VA: American Statistical Association.

Tang and Liu (2010). Sample size recalculation in sequential diagnostic trials, *Biostatistics*, 11, 1, 151–163)

Tang, Emerson, and Zhou (2008). Nonparametric and Semiparametric Group Sequential Methods for Comparing Accuracy of Diagnostic Tests, *Biometrics* 64, 1137–1145

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.

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.

Appendix

IDE Statistical Quality Review Assessment ("Checklist") for Diagnostic Submissions

I. Introduction/Background

1) Is the background of target condition supplied?
2) Is the device clearly described?
3) Is the intended use clearly specified, including target condition, intended subjects (e.g., those with target conditions, asymptomatic subjects), users, setting (single lab, CLIA certified labs etc.), anatomic location or source of lesion, specimen etc., specimen type, platform, and how the device is used?
4) Have the regulatory history and previous submission numbers been provided?

II. Pre-Clinical and Other Studies

Review *of any applicable* pre-clinical and other supporting studies (design and results as appropriate), including the following:

1) Studies to develop or "train" the device (e.g., for cut-offs, or classification model development)
2) Precision and/or Bench studies (studies of variability of repeated measurements by the device under different settings – ranging from repeatability to reproducibility)
3) Analytical studies (e.g., linearity, cutoff determination, limit of blank (LoB), limit of detection (LoD), limit of quantitation (LoQ))
4) Animal studies
5) Clinical feasibility studies
6) Any other studies (US/OUS) that will be used to support application.

III. Pivotal Clinical Study Design

1) Are the purpose of study and the study objectives (such as clinical endpoints, diagnostic performance goal(s), or whether the study involves hypotheses or estimation) clearly described?
2) Is the type of study design (e.g., paired, parallel groups, cross-over, one-arm) described?
3) Regarding study sites, are the number, location, and if appropriate, differences in site characteristics (e.g., testing versus collection) clearly described?
4) Is there a description of how study subjects will be selected (representative cohort, enriched, spectrum of disease to be studied), how they will be screened and how they will be selected for enrollment?
5) Is the storage and handling of study specimens, if applicable (e.g., fresh or archived, and for how long) described?
6) Has the clinical reference standard (CRS) (provides the diagnostic "truth") and/or non-reference comparator devices or diagnostic procedures, if applicable, been clearly defined?
7) Is how the qualification of operators/readers will be assessed clearly described? Is a plan training such personnel in use of investigation device specified?
8) Is there a clear description of the randomization or other schemes for assignment of subjects to groups in a parallel group design, or order of measurements?

9) Is the order and timing of measurement by device, CRS, and comparators described, if applicable?
10) For paired and cross-over designs that involve an order, is there a consideration of a washout period and justification?
11) Has the duration of the study, its follow-up schedule if any, study-recommended timing for study procedures, and any other time-related issues (if relevant) been adequately described?
12) Are blinding schemes: for blinding of the device operators and study subjects to the candidate device, clinical reference standard measurement and other devices/comparators, specified? In a parallel group design, can also refer to blinding evaluators of clinical outcome to the group to which a subject was assigned.
13) Is there a plan for assessment of blinding?
14) Has the strategy and rationale used for selecting a clinically relevant threshold for a qualitative device (pre-specified for pivotal study) been provided?Equivocal/borderline results on candidate device, clinical reference standard and comparator device(s), and how they will be handled in the conduct of the study (should be consistent with intended use) are expected to be pre-specified.
15) Is there a plan for identification and reporting of adverse events, if applicable?
16) Is there a plan for missing data prevention and data collection quality monitoring?

IV. Basic Statistical Analysis Plan

1) Are the primary and secondary clinical endpoints or diagnostic performance measures (DPM) clearly defined?
2) Are the hypotheses (null and alternative) or estimation goals associated with endpoints/DPM clearly expressed both verbally and mathematically?
3) Is the study success criteria with respect to endpoints/DPMs clearly defined?
    a) Do endpoints/DPMs, hypotheses and success criteria support intended use?
    b) Do the endpoints facilitate clinical consequence to subjects? (Consequence of false positives and negatives; consequence of measurement error; consequence of interpretation errors).
4) For studies with hypotheses, is there control of overall Type I error probability with acceptable Type II error (either analytical or simulated)?
5) Does the analysis plan for primary and secondary endpoints/DPMs contain the following applicable elements clearly stated?
    a) Statistical test/methodology for hypotheses testing or estimation, detailed list of covariates in models
    b) Specification of analysis populations (intent-to-diagnose, per protocol)
    c) Important subgroup analyses (e.g., gender, severity of disease, patient conditions that can cross-react with target condition) and interactions
    d) Multiplicity handling for multiple endpoints/diagnostic performance measures and subgroup analyses
    e) Plan for assessment of model assumptions
    f) Model checking diagnostics if simulations are done (e.g., MCMC)
6) Is there a plan to assess, avoid or handle potential biases, such as bias in the selection of subjects relative to the intended use population, bias due to lack of verification of the target condition by clinical reference standard in some subjects, bias arising from an imperfect clinical reference standard, and bias arising from

time device result is taken relative to time clinical reference standard result is taken?

7) Is there a clear plan for missing data handling and a plan for sensitivity analysis for the missing data?
8) Is there a plan for reporting invalid test results?

V. Sample size Determination

1) Is the sample size correctly calculated for hypothesis testing or estimation goals, Are the operating characteristics of the statistical procedure (Type I and Type II error probabilities) correctly reported?
2) Are any underlying statistical models, assumptions and estimates clearly identified and justified?
3) Has a maximum study sample size been reported, incorporating loss-to-follow-up/missing data?

VI. Detailed Data Analysis Plan

1) Is there a plan to report accountability of all subjects/specimens?
2) Is there a plan to assess subject demographic and baseline characteristics and, for two parallel group comparisons, the balance of these characteristics? For the latter is there an analysis plan if the two groups are not balanced?
3) Are important subgroups identified and an analysis plan described for the subgroups?
4) Is there a plan for assessment of poolability of baseline covariates and study outcomes across sites (if applicable) and US/OUS subgroups?
5) Is there an analysis plan if data are not poolable across sites/subgroups?
6) Is there a detailed plan for addressing any relevant biases (e.g., selection bias, verification bias, testing environment bias, and diagnostic truth misclassification by imperfect clinical reference standard)?
7) Will the Statistical Analysis Plan be finalized before either any outcome data become available or before the data are unblinded?

VII. Data lock and submission

1) Does the sponsor have a plan for data lock and electronic submission of all pre-clinical and clinical data including subject-level data Links to preferred format of submission of electronic data:

- http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/ucm134508.htm
- http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/ucm136377.htm

Supplement for Adaptive Designs

1) Is a description of the following design features for the adaptive design provided?

a. Aspect of adaptation: early stop for effectiveness, early stop for futility, sample size re-estimation, adaptive selection of subgroup, randomization ratio (e.g., play-the-winner or drop-the-loser), population enrichment, adaptive threshold in device result, adaptive diagnostic signature (classifier).

b. Is the adaptive method well-established?

- **c.** Will information used for an adaptive decision be revealed or blinded to study participants (investigator, subject, sponsor)?
- **d.** What information is unblinded at interim looks? Is information used for an adaptive decision involve device performance (e.g., diagnostic accuracy, device result, clinical reference standard result) or only information ancillary to device performance (e.g., data variability)?
- **e.** Will interim analysis be based on a intermediate endpoint (e.g., an outcome with a shorter follow-up time)? Is the intermediate endpoint justified?
- **f.** Study attributes, e.g., information unit (per subject, per region), timing / number of interim looks
- **g.** Test statistic(s) at each look
- **h.** Decision rule(s) at each look
- **i.** Method of combining information at final data analysis

2) Have the following statistical considerations been addressed?

- **a.** Controlling study-wide type I error rate
- **b.** Sample size estimation; maximum and minimum sample size
- **c.** Statistical bias in estimates of device performance associated with study design adaptations; bias in confidence interval coverage
- **d.** Potential for increased type II error rate (decreased power) for each study performance goal
- **e.** Details of analytic derivations of statistical properties, if appropriate
- **f.** Use of published literature for support
- **g.** Simulations to characterize and quantify level of uncertainty in each adaptation and impact on statistical study properties
  - (a) Simulated type I error rate under a range of reasonable parameter values for null hypotheses (e.g., device result non-informative for presence or absence of condition of interest)
  - (b) Simulated power under a range of clinically possible alternative hypotheses
  - (c) Sample size distribution for adaptive sample size design
  - (d) Comparison of proposed adaptive design vs. fixed design
  - (e) Computer programming code submitted

3) Have the following logistic considerations been addressed?

- **a.** Written charter of DMC, including the reporting structure to the sponsor and Steering Committee
- **b.** Operating procedures, firewalls, written agreements regarding who performs the unblinded analyses and sponsor/CRO involvement in recommendations for adaptations
- **c.** Specification of entities who would remain blinded from the result and decision of interim looks
- **d.** Communication between the sponsor, FDA and DMC regarding the interim results (meeting minutes)
- **e.** Data collection system and query resolution (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

1. Has the following prior information been provided?
    i. Description of studies, if any, being used as prior information (devices, clinical reference standards, primary endpoints, protocols, treatment groups, study populations)
    ii. Level of data available from each study (e.g., summary statistics by study groups or subject -level data)
    iii. Clinical assessment of exchangeability of parameters across prior and current studies (for diagnostic devices, exchangeability assessment is bivariate in that a pair of performance metrics, e.g., sensitivity and specificity, can be considered to investigate trade-off between frequencies of false negatives and false positives)
    iv. Calculation of prior probability of the study claim
    v. Calculation of prior effective sample size, effective number of subjects expected to be borrowed from prior studies/information
    vi. Description of how to calibrate prior studies with proposed study, e.g., using measured covariates
    vii. Description of model used to incorporate prior data (e.g., hierarchical, power prior, commensurate prior)

2. Has information on the following adaptive design issues been provided (if applicable)?
    a. Stopping threshold for enrollment (e.g., predictive probability threshold)
    b. Stopping threshold for study success and for futility
    c. Information measure being monitored at interim stages (e.g., sample size, effective sample size, posterior standard deviation)
    d. Description of predictive model for yet to be observed data, if applicable. Is sensitivity analysis needed to determine if results are robust to model specification (e.g., subjective hyperprior, data missing at random)?
    e. Minimum number of subjects at first interim look (e.g., for safety, effectiveness, futility, stopping enrollment, etc.).

3. Have the following operating characteristics been provided?
    a. (Usually simulated) type I error rate and power, (perhaps assuming various amounts of borrowing from prior studies, if applicable)?
    b. How was the type I error rate calculated? (e.g., at the point null, averaged over prior null distribution, prior to the prior information)

4. Are the following modeling issues adequately addressed?
    a. Is the posterior proper? (May not be if prior is improper.)
    b. Will the analysis use Markov Chain Monte Carlo (MCMC) simulation to sample from posterior distribution? If so, will the MCMC results be checked for convergence?
    c. Are the parameters identifiable?
    d. Is the prior used for a hyperparameter subjective? Informative? Should sensitivity to the hyperprior be investigated?
    e. Will the model be checked for adequate fit by the data? (e.g., choose a statistic related to a model assumption, compare its observed value to its predictive distribution under the model, summarize fit by the Bayesian p-value, the tail area of the predictive distribution at the observed value)