Millennium Pharmaceuticals, Inc.
75 Sidney Street
Cambridge, MA 02139
Tel 617 679 7000
www.millennium.com

20 June 2003

Docket 03D-0120

Docket Management Branch
Division of Management Systems and Policy
Office of Human Resources and Management Services
Food and Drug Administration
5630 Fishers Lane
Rm. 1061 HFA-305
Rockville, MD 20852

Millennium Pharmaceuticals, Inc. Commentary on the Draft Guidance, *"Multiplex Tests for Heritable DNA Markers, Mutations and Expression Patterns"*

Dear Sir or Madam,

Millennium Pharmaceuticals, Inc. ("Millennium") is a global research-based biopharmaceutical company and leader in genomic drug discovery based in Cambridge, Massachusetts with a European affiliate in London, UK. Millennium's research, development and commercialization activities are focused on genomic approaches to the innovation of breakthrough products to treat cancer and endocrine, metabolic, cardiovascular and inflammatory diseases.

Millennium is grateful for the opportunity to provide comments for consideration on this important draft guidance.

General Comments

Collaboration between CDRH, CDER and CBER

The draft guidance is specifically intended for the submission of a diagnostic test to CDRH for pre-marketing approval and it remains unclear whether this guidance was the result of a collaborative effort between CDRH, CDER and CBER. Multiplex tests may also be used during clinical development and the regulatory hurdles for such applications may differ from those for a marketable device. The guidance, as written, does not take this into consideration and should distinguish between these two scenarios. CDRH, CDER and CBER should, at the very least, agree to common analytical validation standards, even if the Centers use separate clinical evaluation criteria. We would hope that a future draft of the guidance would indicate the extent of common acceptance of standards between the Centers and the specific points of difference.

C3

03D-0120

Continued .../

Consistency throughout document

The guidance contains a fairly high level overview on some topics and great detail on other topics. We suggest that there should be a more consistent approach throughout the document for all topics discussed.

Intended uses

The guidance would be clearer and easier to follow if it was divided into sections based on intended uses. Our recommendation is to divide the guidance into 1) genetic tests and associated dichotomous and trichotomous measurements, and 2) gene expression tests and continuous measurements.

Separation of data

If a device, such as a microarray, has the capacity to measure 30,000 or more covariates, and only 50 genes or genotypes present on the device constitute the test, what happens to the other variables that are measured? Will a separate test have to be developed in order to incorporate only those genes/genotypes deemed significant? If so, must a diagnostic test that will be cited in the label of a drug be used in the drug's safety and efficacy trials, or can bridging studies be done, as for the approval of HercepTest® ? These issues should be addressed in the guidance.

Specific Comments

Introduction

Further clarification is needed on the specific types of tests this document is intended to cover. To this end, the revision of the draft should provide a more detailed description of a multiplex test. While microarrays meet this definition, it is not clear whether multiple analytes, assessed in independent assays but combined into a predictive test, would also meet the definition. We believe that it should.

One can imagine many different ways of assessing DNA markers without using arrays, but using procedures that measure genotype at one locus or many loci. The simplest case to consider is the assessment of genotype at one marker. This could be technically construed as a multiplex test, because two alleles are being assayed. More complex tests include assessing multiple markers (or mutations) within a gene, assessing the haplotypes of a gene directly, estimating a haplotype from genotype data, and combining genotype data or haplotype data across many genes. The complexity of multiplex tests goes beyond these statements made for genetics. The same type of queries could be made for multiple quantitative rt-PCR or ELISA tests. The major point here is that the FDA should be much more specific about the definition of "multiplex".

The purpose and meaning of the following statement should be made clear: *"Clinical studies should account for disease prevalence in the populations studied."*

Recommendations for the Preparation of the Multiplex Test Application

We suggest rewriting the analytical validation section to reflect a general structure in which guidance for assay components, including arrays, reagents, samples, and equipment, and software is dealt with in Part A (see rewrite below) and the actual validation of the specific assay and

platform at a site (Pharma or CRO) is described in Part B. This section would include assay validation experiments and specific documentation which are not generic to the components and reflect the utilization of all the components described in Part A to address a specific assay, used to derive data associated with a clinical study, test plan, or operating lab.

- Certain sections of the draft guidance are written for single analyte tests specifically and are not applicable - see the following rewrite of the analytical validation section for removal of certain text.
- The assay validation section should be expanded to specifically recommend assay validation activities such as documentation, site specific validation, tracking of process controls, and assay performance over time.
- In general, the language needs to be clarified for multiplex tests with continuous measurements per analyte. Specifically, the section referring to assay sensitivity and reproducibility should reflect the use of these terms in a multiplex context, i.e. sensitivity for all genes or a subset.
- This section may have to be written differently for multiplex genetic tests vs. gene expression assays.

II. Analytical Validation

Proposed New Text below in regular font, existing guidance text in *italics*.:

A. Validation or Assay Components/Systems

We recommend that submissions include analytical data that demonstrate that the device performs accurately and reliably under given conditions; this may include:

1. Array manufacture - *Product design, manufacturing, and controls must conform with applicable parts of the Quality System regulation (QSR) as set forth in 21 CFR §820, see also 21 CFR § 814.20(b)(4)(v).*

Specifically, the following elements of arrays and multiplex platforms should be well-characterized: design, internal controls used, oligonucleotides, primers, probes, or other capture elements, conditions for producing arrays, including washing procedure and drying conditions (e.g., temperature, length of time), methods used to attach the target material to the matrix, composition and spatial layout of arrays or other spatially fixed platforms, specificity for markers or targets, and stability of the platform. In addition, annotation (gene ID, etc.) associated with each nucleic sequence should be described in its derivation and source.

2. *Specimen/sample - (for each claimed matrix): identity, preparation, acceptance criteria where applicable, and methods for determining label incorporation, probe length, and so forth, for samples that will be hybridized to the array. Also, include specimen collection, storage, and handling conditions.* A common nomenclature should be adopted for reference to probe and target.

3. Reagent Assay components - *including buffers, enzymes, signal detection systems such as fluorescent dyes, chemiluminescent reagents, other signaling reagents,* controls and/or calibrators: negative and positive controls, characterized as internal or external.

4. Instrumentation -

a. Characterization: Characterize instruments used in the assay, including how the instrument assigns values to or interprets assay variables such as feature location, size, concentration, volume, drying of small samples, effect on small volume reactions and its impact on test results.

b. Calibration: Describe instrument calibration.

c. Uncertainties: Describe sources and estimates of uncertainties in results introduced by hardware components.

5. Software/LIMS – Design must conform with applicable parts of the Quality System regulation (QSR) as set forth in 21 CFR Part 11.

B. Assay Validation/Validation of Specific Performance Characteristics: Analytical Laboratory Studies

1. *We recommend that you describe the following performance characteristics* associated with the assay platform as operated at the site of use *for each target, pattern, marker or mutation claimed in the intended use statement]:*

a. *Assay sensitivity: ability to accurately identify positive samples.* [Comment - It is not clear what "positive sample" means here for expression measuring devices. For genetic tests, a positive sample means a disease allele is present. For expression tests, it is unclear if this means that "absolute expression level" is above some lower bound. Two-channel cDNA microarrays are more accurate on relative expression level than absolute expression level. FDA should clarify if this means that the relative expression ratio to a "designed" reference sample is above a lower bound.]

b. *Reproducibility: Consult NCCLS EP-5A and EP-12A for information on reproducibility studies, http://www.nccls.org/.*

c. *Validation of controls (samples and array features), definition of quality control metrics.*

d. *Assay range.*

e. *Effect of excess sample and limiting sample.* High and Low limits (i.e. amount of RNA sampled, amount of cRNA hybridized) *Investigate the sample concentrations and conditions that reproducibly yield acceptable results*

f. *Assay specificity and interfering substances (endogenous and exogenous).*

2. Array processing

We recommend that the following are described:

1. Performance of assay validation experiments and documentation

2. *Optimization of multiple simultaneous target detection/differentiation, for example, hybridization conditions, concentration of reactants, control of specificity.*

3. *Potential for sample carryover* during applicable processes

4. *Limiting factors of the array, including saturation level of hybridization.*

5. All QC metrics and Quality Assurance measures

3. Data transformation and processing

Computational methods for data processing specific to the assay and platform. *We recommend that you develop computational methods using the CDRH software development and validation guidance documents that are available at* http://www.fda.gov/search/databases.html to address conversion or raw data (images, signal measurements) to expression values, genotypes. This can include systems for the following:

1. Image processing
2. Data condensation
3. Algorithms for conversion to expression values, genotypes
4. Generation of noise/process variation metrics, p-values

End of Proposed New Text

III. Comparison studies using clinical samples

Comment: A guidance is also needed to describe how clinical samples should be collected.

The guidance states:

A. *Comparison to another device: Results of comparison studies with another well-characterized or predicate device; usually reported as percent agreement.*

The use of multiplex tests may make it difficult to compare to another device in order to establish performance. Non-multiplex devices may not be available. Also, percent agreement between two devices could potentially mean that both tests make the same mistakes.

B. Comparison to a Reference Method: Results of comparison studies to a validated reference method or clinical diagnosis; usually reported as sensitivity and specificity.

1) A receiver operating characteristic (ROC) curve would be preferable to a report of sensitivity and specificity.
2) It is unlikely that comparison to a validated reference method will be useful. The critical comparison is to independent clinical samples in a validation setting.

D. Identification of analytical/technical false positive or false negative results, estimates of expected assay failure rates.

Technical false positive or false negative rates may not be separately attainable from clinical comparison, especially in the absence of an alternative diagnostic method. The analytical validation section should be used to guide expected analytical error rates.

IV. Clinical Evaluation Studies Comparing Test Performance to Accepted Diagnostic Procedure(s)

Clarification is needed of the potential safety issues that would be associated with an *in vitro* diagnostic device.

The guidance refers to the definition of "Clinical Truth" in evaluating clinical performance. One expects that this pertains to the comparison studies described in Section III as well. "Clinical Truth" may be a troublesome standard, because we believe it unlikely that any independent diagnostic assessment may be available in most cases. In the absence of "true" clinical truth, assessment of diagnostic performance against independent clinical samples may be the best way to evaluate performance. In the absence of long-term patient outcome data, unbiased surrogate outcome data may be a proposed acceptable standard for patient outcome.

The guidance states:

3. Clinical data: Validate expression patterns, genotype/phenotype correlations, and so on, on a statistically adequate number of specimens for each intended use, including clinical samples for all matrices claimed in the intended use statement; verify with a second detection system (e.g. quantitative RT-PCR) if applicable.

- A description of the standards against which expression patterns should be validated is required.
- Further clarification is needed for the statement "*including clinical samples for all matrices claimed in the intended use statement.*"
- This guidance should address what samples will need to be verified by a second detection system. Will all samples need to be verified, or just a subset? Sample quantities may be a limiting factor.

The guidance states:

When defining the populations used, submissions should include the following information::

- *Number of samples from the normal population with samples summarized according to appropriate demographic characteristics.*

The following questions need to be addressed in this section:

- What is meant by "demographic characteristics"?
- What defines "normal" from a genetic/genomic perspective?
- Will normals be required if comparisons will only be made within a disease group (e.g. a comparison of responders to non-responders)?

The guidance states:

5. Statistical method: Describe statistical methods used for calculations. Measures of precision, e.g., confidence intervals, should be described and presented.

The guidance should address how the validity of analysis will be handled when multiple statistical methods are available and there is no consensus on which method is superior and/or there are no predicates.

Appendix I: General considerations for planning and evaluating clinical studies.

If these general recommendations are different from those given in other guidance documents, then it would be helpful to cross-reference those documents and/or point out the differences in this document. In addition, more guidance will be needed to address the question of when retrospective studies are acceptable.

The guidance states:

4. Determine sample size prior to beginning the clinical study. The sample size should have sufficient statistical power to detect differences of clinical importance for each marker, mutation, or pattern. FDA will consider alternate data sets in cases with a small available sample size, for example, a disease or condition having a low prevalence or with markers or mutations of very low frequency.

The guidance should define what would be considered an alternative data set, how such a data set will be powered and how it will be evaluated.

The guidance states:

7. Display genotype data in the appropriate N x N table (e.g., 3 X 3 for homozygous wild-type, heterozygous and homozygous mutation) where applicable.

While this summary may be useful, it seems odd that genotypic data are singled out. Other types of summaries could be made for various genomic/proteomic markers.

The guidance states:

9. Include samples from individuals with diseases or conditions that may cause false positive or false negative results with the device (i.e., within the differential diagnosis), if appropriate.

In order to fulfill this requirement, control samples should be available, such as from a repository. Guidance on how to provide these control samples should be included.

This section is missing points #10-12.

The guidance states:

14. Perform studies using appropriate methods for quality control. Describe the materials and methods used to assess quality control.

It needs to be clarified whether the studies must be conducted under GLP conditions. What if GLP or CLIA certified laboratories do not exist?

The guidance states:

15. Describe how the cut-off point (often the distinction between positive and negative, or the medical decision limit) was determined, if appropriate. Describe the performance characteristics the cutoff identifies for each marker/mutation. The description of how each cut-off was determined should include the statistical method used (e.g., receiver operating characteristic curve).

For clarity, we recommend that this section should be broken out into "DNA" versus "expression" profiling.

Appendix II: Statistical considerations for analyzing array data

This appendix seems to be a compilation of helpful hints for analysis, but it is far from comprehensive with respect to methods that could be applied to multiplex data. For example, there is no description of haplotyping methods. If this appendix originally comes from other material, the original document should be cited and this section should stress how multiplex tests may be different.

The method for comparison between studies is not clear. If there is no measure of truth, why does the measure of a new method have to be in agreement with the measure of a control method? In the "*Use of null hypothesis testing*" paragraph, the use of the slope of linear regression to test agreement (being close to 1) is suggested. Even if the slope equals 1, it still could mean both methods are not giving the correct value. If the new method measures the expression level lower than the control method in a consistent way such that the classification with a proper cutoff value is as good as the control method, does this mean the two methods are in agreement?

Without a "true" measurement, how can reproducibility be evaluated? If we knew the true expression level of a gene, for example, between 1 and 1.2, we would like to have the mean of the expression level between 1 and 1.2 at every site (for site-to-site reproducibility) with a small standard deviation and small number (or none) of individual measurements outside of (1, 1.2). Similar comments apply to array-to-array reproducibility, day-to-day reproducibility, and sample-to-sample reproducibility, and so on.

For a diagnostic device, single sensitivity/specificity or positive or negative predicted value is not sufficient, because a device is usually not used under a single point of ROC. ROC should be provided for the full range, which the device is intended to use. Without an ROC curve, bias may occur since investigators can choose a best point of sensitivity and specificity for their device for the comparison. Sampling plan and acceptance criteria for performance characteristics for analytical validation should be provided. The FDA draft guidance recommends NCCLS guidelines be consulted. This, however, seems to be inconsistent with the recommendations given in other guidances for analytical validation, which requires the performance characteristics such as accuracy, precision, linearity, range, LOD, LOQ, and ruggedness as described in the USP/NF be considered.

It is suggested that qualitative and quantitative analytical validation be distinguished in the draft guidance. Statistical methods for assessment of qualitative and quantitative analytical validation could be very different under different study designs.

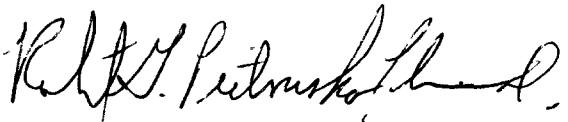
The draft guidance does not provide information regarding what sample sizes are required for specific studies.

The leave-one-out method is known to be biased, and should not be recommended. A better approach in the absence of a completely independent dataset would be multiply repeated X-fold cross-validation (where x = 5 or 10).

As indicated in the guidance, FDA may request different types of data and statistical analyses in pre-market applications for *in vitro* diagnostic tests. However, little information as to what types of data and statistical analyses that may be required are provided.

Variation associated with multiplex tests is always a concern, which may have an impact on the accuracy and reliability of the test results. The draft guidance does not address the issue of variability that may come from disparate sources.

Sincerely,



Robert G. Pietrusko, Pharm.D.
Vice-President
Worldwide Regulatory Affairs and Pharmacovigilance
Millennium Pharmaceuticals, Inc.