Genentech
A Member of the Roche Group

FDA Draft Guidance for Industry: Statistical Approaches to Evaluate Analytical Similarity

Docket No. FDA-2017-D-5525:

Comments from Genentech a Member of the Roche Group

Dear Sir/Madam:

Genentech, a member of the Roche Group, thanks the Food and Drug Administration (FDA) for the opportunity to submit comments on the Draft Guidance on Statistical Approaches to Evaluate Analytical Similarity.

Genentech applauds the FDA and its statisticians for recognizing the importance of statistics in evaluating analytical similarity of a therapeutic protein product to a reference product. The appropriateness of the statistical methods will determine the value of the analytical similarity assessment in evaluating the biological product for licensing under section 351(k) of the Public Health Service Act (42 U.S.C. 262(k)). We ask the Agency to consider our general and specific comments during the review and finalization of the Draft Guidance.

We would be pleased to provide further input or clarification of our comments, as needed.

Sincerely,

Eric J. Olson
Vice President, U.S. Product Development Regulatory
Genentech, a Member of the Roche Group
1 DNA Way, South San Francisco, CA 94080
Email: olson@gene.com, Phone: 650-467-7711

Daniel A. Coleman, Ph.D.
Associate Director, Nonclinical Biostatistics
Genentech, 1 DNA Way, South San Francisco, 94080
Email: dcole@gene.com, Phone: 650-703-2727

General Comments

A rigorous statistical assessment of analytical similarity depends on appropriate statistical methods combined with good experimental design. The best way to ensure rigor is to have both the statistical methods and the experimental design described prospectively in a protocol, not a plan. In particular, the protocol will describe the data collected prior to experimentation on the reference and biosimilar products, the data to be collected during experimentation, the statistical methods and acceptance criteria to determine if the data collected during experimentation supports the hypothesis of similarity. One of the guidance's first sections should describe the protocol, the interactions the Agency is prepared to have with the sponsor on protocol development, and the milestones in place prior to performing the assessment, e.g., process lock, assay validations, stability determination. Finally, the proposed statistical methods, e.g., equivalence testing, should be connected to their experimental designs; please see our specific comments (Line 211).

The draft guidance notes important differences between the development of biosimilars for orphan drugs and for non-orphan drugs. From a statistical perspective the key difference is sample sizes available to make a statistical assessment. Minimum sample sizes affect the available statistical approaches. Please consider limiting the scope of this document to non-orphan drugs, removing all references to orphan drugs, and stressing the minimum sample sizes needed for the assessment. Another document could perhaps be written for orphan products.

There is a misconception that testing of multiple attributes will likely lead to the incorrect classification of truly-biosimilar products as not biosimilar. This is illustrated by recent comments of Martin Schiestl[1], chief science officer at Sandoz, at DIA, Oct 25, 2017. Multiple testing is well studied in the statistics literature[2] and has been successfully employed in clinical trials for many years. In clinical trials, multiple testing procedures are designed to control consumer risk; in analytical similarity testing, the procedures need to be designed to control producer risk as well. Please see our specific comments (Line 82-86). We urge the authors to promote the appropriate use of multiple testing procedures to control both types of risk.

1. Links to news articles discussing Martin Schiestl comments:

http://www.raps.org/Regulatory-Focus/News/2017/10/24/28737/Sandoz-Raises-Questions-With-FDA-Draft-Guidance-on-Statistical-Approaches-for-Biosimilars/

https://endpts.com/sandoz-raises-questions-with-fda-draft-guidance-on-statistical-approaches-for-biosimilars/

2. Books on multiple testing:

Multiple Testing Problems in Pharmaceutical Statistics, Edited by Alex Dmitrienko, Ajit C. Tamhane, Frank Bretz, Series Editor: Shein-Chung Chow, Chapman and Hall/CRC, 2009

Multiple Comparisons: Theory and Methods, Jason Hsu, Chapman and Hall/CRC, 1996

Specific Comments

Title

Comment: The scope of this document does not include process changes or process transfers made by the innovator. Other documents addressing biosimilars have titles that clearly indicate scope e.g.: "Demonstrating Biosimilarity to a Reference Product" and "Quality Considerations in Demonstrating Biosimilarity of a Therapeutic Protein Product to a Reference Product".
Proposed Change: Consider renaming the document to include the scope, for example: "Statistical Approaches for Evaluating Analytical Similarity in Demonstrating Biosimilarity to a Reference Product" or "Statistical Approaches for Evaluating Analytical Similarity in Demonstrating Biosimilarity to a Non-Orphan Reference Product".

Line 74-76

Comment: An analytical similarity assessment made late in the biosimilar development process will bring more value to the totality of evidence. For example, if the assessment is made after critical development milestones, e.g., lock down of the commercial biosimilar process, assay qualifications/validations, stability determinations, the assessment will be more relevant to the actual biosimilar manufacturing process and therefore more valuable in determining if the proposed product is biosimilar. This guidance can promote more relevant assessments by clarifying the Agency's expectations about the timing of the assessment and the Agency's role in plan review and development.
Proposed Change: Please consider discussing the Agency's expectation for the placement of the assessment in the biosimilar development process and the Agency's role in the plan. Consider moving the sentence (Line 326) "The analytical similarity assessment plan should be discussed with the Agency as early in the biosimilar development program as possible so that agreement can be reached on which attributes/assays should be evaluated in each tier" to this section. Also please describe how this assessment fits into the control system and continuous process validation (CPV) of the biosimilar manufacturing process.

Line 82-86

Comment: Multiple testing has been extensively studied in the statistical literature and has been successfully used by our colleagues in clinical statistics. A variety of statistical approaches are used in clinical trials with multiple primary endpoints. These approaches control both type 1 and type 2 errors of the primary endpoints through experimental design and sample size. The precision of the estimates of the secondary endpoints are typically not controlled. In the proposed analytical assessment approach attributes are assigned to tier 1 and tier 2 for testing based on risk to patients. This is analogous to clinical trials classification of endpoints to primary and secondary based on clinical relevance.
Proposed Change: Please consider removing the following text: "Third, there are a large number of potential quality attributes that can be compared in an evaluation of analytical similarity, and subjecting all of these attributes to formal statistical tests in the context of limited lots could lead to concluding incorrectly that a large number of truly highly-similar products are not highly similar." The text does not recognize the feasibility of multiple testing and may be interpreted as permission to jettison rigorous statistics and substitute non-statistical acceptance testing, e.g., the quality range method (QRM).
Please consider consolidating the discussion of a false negative (rejecting the biosimilar when it is truly similar) into one subsection.
We recommend that the Agency recognize that the chance of passing all tests in the assessment should be high, e.g., 80 or 90%, under "***sameness***", i.e., when the biosimilar product attributes have the same distributions as the reference product attributes.

We recommend the following approach motivated by clinical trials for using equivalence testing in tiers 1 and 2. Tier 1 consists of tests of 1, 2, or 3 "high" risk attributes; the margins for these tests are determined by scientific knowledge, clinical impact, or 1.5 times the reference standard deviation. Tier 1 determines the sample size; the sample size is adjusted to provide a high chance of passing under sameness. Tier 2 consists of tests of 5 to 10 "low" risk attributes, since the sample size is fixed in tier 1, the margins are adjusted to provide a high chance of passing under sameness, e.g., a constant k is determined so that margins (k times reference standard deviations) are set to provide a high chance of passing under sameness. Here k will be much larger than the recommended 1.5 used in tier 1.

Line127-129
Comment: Assessing the analytical similarity of an attribute that changes over its shelf life presents significant challenges. Understanding the attribute's degradation rate on the reference material would enable the estimation of a sample's attribute at drug product release and the average attribute value of the reference material at drug product release.
Proposed Change: Please consider adding a subsection dedicated to assessing the analytical similarity of an attribute that changes over its shelf life. What information should the sponsor have regarding stability of the reference going into the assessment? Will a model be needed to infer the attribute release values in the reference samples or are other options available? Does the guidance offer any insights into how to do regression or other assessments that estimate release test results, and does this add uncertainty that needs to be overcome with additional sampling?

Line 169
Comment: The text "Methods of varying statistical rigor should be applied depending on the risk ranking of the quality attributes" implies that statistical rigor is measured on a continuum. Equivalence testing is statistically rigorous, but ad-hoc procedures, e.g., the quality range method (QRM), are not. Incorporation of a statistical method into an ad-hoc procedure, e.g., using tolerance intervals to set the range in the QRM, does not make the ad-hoc procedure statistically rigorous. Regardless, if an attribute is determined to be low risk, the margins in an equivalence test could be set wider or other adjustments could be made to reduce producer's risk.
Proposed Change: Please consider removing the text.

Line 195
Comment: The words "poor" and "high" in the text, "Poor assay performance, including high assay variability, should not be used to justify selection of either a particular evaluation tier or an inappropriately broad similarity acceptance criteria." are not defined. A more direct approach to accomplish the objective of the text is to address the qualification/validation status of the assays used in the assessment. For example, validation of assays per the sponsor's SOPs following the ICH Q2B Guideline would set "fit-for-purpose" acceptance criteria on the assays.
Proposed Change: We recommend that the Agency describes their position on the status of assays used in the assessment, e.g., qualification/validation. The description should use assay terminology defined in other guidance documents, e.g., ICH Q2B. The requirements should be similar to what would be expected in the BLA of the innovator's product.

Line 204
Comment: The analytical similarity assessment plan is a key component in licensing of the biosimilar. The integrity of the assessment relies on pre-specification of the assessment plans activities. The assessment plan should be treated with the same rigor as a protocol. As a protocol its lifecycle, e.g., preparation, approval, amendments, archiving, would be regulated by the sponsor's SOP governing protocols.

Proposed Change: Ideally, the analytical similarity assessment plan should be a protocol and renamed to "Analytical Similarity Assessment Protocol".

Line 211

Comment: Statistical rigor requires prospectively described experimental design and analysis. A key aspect of the experimental design is a description of the data to be collected during the experiment. Data collected during the experiment is random and can be used for statistical inference but not for acceptance criteria, e.g. equivalence margins. Data collected prior to experimentation is not random and cannot be used for statistical inference but can be used to set prospective acceptance criteria.

Consider two experimental designs:

1) A representative collection of reference samples is used to prospectively set a target for the biosimilar process and equivalence margins. Only biosimilar data is collected during the experiment. The analysis is an equivalence test consisting of two 1-sample tests.

2) A representative collection of reference samples is used to prospectively set equivalence margins. Both reference and biosimilar data is collected during the experiment. The analysis is an equivalence test consisting of two 2-sample tests.

An important advantage of 1) is that the sponsor can perform the assessment at lab scale, if similarity is declared the sponsor should have confidence in passing at manufacturing scale.

Proposed Change:

We suggest that the assessment plan stage "Determination of the statistical methods to be used for evaluating each quality attribute based on the risk ranking and on other factors" be generalized to "Determination of Experimental Design and Statistical Analysis" and that the data to be collected during the experiment as well as all data collected prior to plan approval be described along with their uses. The Agency's perspective on the two experimental designs described above would be valuable to sponsors.

Line 316

Comment: Biosimilar development typically starts with the collection of reference samples and the development of assays to measure attributes of reference samples and biosimilar product.

Proposed Change: Please consider changing the text "In some cases, it may be necessary to first collect preliminary data (e.g., to get an initial estimate of the variability of the reference product's attribute or to select an assay at the outset before finalizing the statistical analysis plan)" to "In most cases, the biosimilar development process will begin with the collection of reference samples to develop and validate assays, understand the distributions of the reference product's attributes, develop a biosimilar manufacturing process, and compare biosimilar product from small scale manufacturing to reference product. This same data may be used in the analytical similarity assessment plan to prospectively set targets and/or margins for equivalence testing."

Line 342

Comment: Statistical tests require random data, i.e., data specified in the analytical similarity assessment plan and collected during experimentation. Data collected prior to the assessment plan is not random.

Proposed Change: Consider adding a discussion before Line 342 outlining what data is random. Please clarify that a 2-sample test is used when both reference and biosimilar data are random and that a 1-sample test is used when the biosimilar data is random and the reference data is not.

Line 410-414:
Comment: The QRM has been accepted by the Agency as a due diligence to demonstrate process comparability after a process change or transfer. Its objectivity relies on pre-specification of the range not on the method for determining the range. It is our understanding that the QRM has origins in pharmaceutical chemistry, not statistics. It is certainly not a statistical inferential method and without a specific method for setting the range, the operating characteristics of the QRM are undetermined.
Proposed Change: Please consider replacing the QRM for each tier 2 attribute with an equivalence test that has a margin set wide enough so that there is a high probability of passing all attributes under the assumption of sameness. We recognize the margins will be wide, but the procedure is statistically rigorous and may be the best statistical procedure available given the circumstances. If the Agency continues to accept the QRM for tier 2 testing, please consider stating the acceptable methods for setting the range, so that the operating characteristics can be determined and evaluated.