Developing Analytical Standards for NGS Testing

The information and questions contained in this document are not binding and do not create or propose new requirements or expectations for affected parties, nor is this document meant to convey FDA's proposed or recommended approaches or guidance. Rather, the information contained in this document offers background and considerations regarding analytical standards for NGS for discussion at FDA's public workshop on November 12, 2015.

Goal

As part of the President's Precision Medicine Initiative (PMI), FDA is considering novel ways to optimize its regulation of Next Generation Sequencing (NGS) tests for human genomes. The ultimate goal of this effort is to develop a flexible, adaptive regulatory approach that ensures that patients receive accurate and meaningful results, while accommodating innovation in test development. FDA posted a paper in December 2014 discussing possible strategies it is considering to accomplish this goal, and obtained stakeholder feedback in a public workshop held on February 20, 2015.

In brief, these strategies involve: 1) identifying and implementing analytical standards that would ensure that NGS tests produce accurate and reliable results; and 2) developing ways to use well-curated databases of genetic variants to guide clinical interpretation of NGS test results. After analysis of public feedback, FDA has further developed more specific concepts for the analytical and clinical strategies. The topic of clinical interpretation is discussed in a companion paper[1] while this paper discusses standards-based strategies to assure that NGS tests produce accurate and reliable analytical results. Identifying such strategies is a critical step in developing a novel approach toward regulating NGS tests. The concepts developed in this paper will be discussed in a public workshop on November 12, 2015;[2] interested parties may provide comment at that time, or submit written comments to an open docket.[3]

Scope

This paper describes two different approaches that might be used to assure the development of NGS tests that are analytically valid. These approaches represent ends of a spectrum from predefining individual *performance standards* for each NGS test, i.e., establishing specific metrics and acceptance criteria that the test would have to satisfy, to the development of *design concept standards* that would ensure, when complied with, that the manufacturer can achieve the proper design and validation of an NGS test. Below, we provide an outline of the structure of these two types of standards. FDA is requesting public input on whether performance standards, design concept standards, a combination of the two, or another standard would best provide assurances that all NGS tests can provide accurate and

[1] http://www.fda.gov/downloads/MedicalDevices/NewsEvents/WorkshopsConferences/UCM467421.pdf
[2] http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm459450.htm
[3] http://www.regulations.gov/#!docketDetail;rpp=100;so=DESC;sb=docId;po=0;D=FDA-2015-N-3015

reliable analytical results. FDA will use this input in drafting future proposals on the regulation of NGS tests. *The concepts below are presented for the purposes of public discussion, and are not statements of current or proposed regulatory policy.*

Background

The sequencing of the human genome and the subsequent increase in our understanding of the relationship between specific genetic variants and disease has already begun to improve health through identification of novel disease markers, and development of tailored treatments and prevention strategies based on the specific characteristics of each individual – Precision Medicine. Next generation sequencing (NGS) technologies have significantly expanded our ability to derive this comprehensive genetic information on individuals in a relatively rapid manner. In order to achieve the full potential of precision medicine, technologies such as NGS must produce accurate, reproducible, and meaningful results relevant to a person's medical condition.

As part of the President's Precision Medicine Initiative (PMI), the FDA is developing new regulatory strategies for NGS tests that foster innovation in test development, while ensuring that the data produced by these tests are accurate and reliable. In a February 2015 workshop[4] and a previously published discussion paper,[5] FDA discussed the possibility of developing a standards-based approach to achieve this goal. Compliance with standards could substitute for premarket clearance or approval by FDA of each individual test.

Traditionally, FDA assesses the safety and effectiveness of a given test by reviewing the analytical and clinical performance of the test in a premarket submission. To assess analytical validity, FDA evaluates the specific performance characteristics of each test based on its intended use, including specificity, sensitivity, positive percent agreement, negative percent agreement, precision, and other relevant metrics. When a test detects multiple analytes, FDA then reviews performance data for each of those analytes. Marketing authorization fundamentally relies on the calculus of whether the benefits of the test outweigh its risks.[6] FDA has not in general predefined specific performance targets that must be met for a test, such as a predefined level of accuracy, although in some cases it has done so in special controls.

NGS-based tests have the capacity to produce, in a single test, data for up to billions of individual analytes. This large number of analytes, and even larger number of possible results, makes it infeasible for test developers to provide and FDA to review performance data for each analyte. Therefore, FDA has

---

[4] http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm427296.htm
[5] http://www.fda.gov/downloads/MedicalDevices/NewsEvents/WorkshopsConferences/UCM427869.pdf
[6] For devices subject to 510(k), this calculus is conducted when the device type is classified. Review of a 510(k) determines whether a device is substantially equivalent to a predicate device. Substantial equivalence review compares the new device to the predicate in terms of intended use and technological characteristics, including performance.

previously used a flexible regulatory approach to address this issue. In clearing the Illumina MiSeqDx NGS platform,[7] FDA determined that performance of the device in detecting representative variants with different properties and from different genomic regions represented a reasonable demonstration of the device's overall performance, which formed the basis of marketing authorization. This approach allowed FDA to infer performance across the entire genome rather than requiring the device manufacturer to submit data to support the analytical performance on each possible variant the instrument could detect.

FDA's consideration of a standards-based approach to provide oversight of NGS tests is based on the Agency's recognition that the number of analytes queried and the large volume of data produced call for new strategies to ensure that analytical output for a given test is of sufficiently high quality to inform correct clinical interpretations. Public comment on the December 2014 discussion paper and at the February 2015 workshop was largely in favor of a standards-based approach to regulatory oversight of NGS-based tests, but recognized some challenges in developing these standards. Based on these comments, FDA is considering and seeks comment on potential approaches to oversight of these tests – documentary performance standards, design concept standards, or other approaches – that it may consider as forming the core of a standards-based approach. FDA wishes to understand whether either standard approach, some combination thereof, or another potential option could be used to assure the analytical performance of NGS tests.

Potential Approaches to Ensuring Analytical Validity

Analytical validation is a process that is intended to evaluate the measurement/detection performance of a test across all relevant metrics, and is conducted to assure that measurements made by the tests are accurate and reliable. NGS analytical validation may be tailored toward detection of known changes in specific genes, and detection of known and novel changes in genes, exomes, and genomes. Although analytical validation strategies for similar tests, e.g., targeted mutation tests, will often have similar types of studies, the performance characteristics of tests for different analytes may differ markedly based on what is technologically possible given a particular sample type and target(s) as well as on the level of performance needed to demonstrate that the test can achieve its intended use when used according to defined procedures.

Standards based approaches to analytical validation

NGS is highly flexible, rapidly changing technology that can provide the basis for many tests with many different intended uses. In considering the most appropriate approach to oversight, FDA has identified approaches on both ends of the spectrum of flexibility and specificity for evaluation of analytical performance which are described below. A more flexible, but less specific approach is to implement development design concept standards that would rely on the established ability of the developer to

[7] http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?ID=DEN130011

implement well-described principles of test design and validation, and thus consistently generate tests that meet user needs, with performance characteristics that allow correct clinical interpretation. A less flexible, but more specific approach is to define, for a developed test, individual performance evaluation studies that must be conducted and specific performance criteria that must be met. There are merits and drawbacks to each of these approaches, in that the first relies heavily on assessment that developers know how to successfully develop quality tests, but does not necessarily involve FDA premarket review of each test developed and validated, and the second provides concrete metrics and performance specifications applied after the test has been developed, but may allow less flexibility to accommodate changes in technology.

Design Concept Standards

For purposes of this document, FDA defines a design concept standard (in the context of an NGS test) as a set of defined activities and goals that are documented, and when complied with, are expected to yield a product that has the intended characteristics and consistently delivers results within the established acceptance criteria.

A design concept standard would not specify the design of any particular NGS test; instead, the use of well-described principles of design would enable the conceptualization of a test from beginning to end. After defining in specific terms the intended use of the test that is desired (specified by user requirements), developers would determine the specific test components, the requisites of each, and the impact of each requisite on the other design elements. Potential components of an NGS test might include sample collection and processing, sample and library preparation, sequence generation/base calling, mapping and alignment, and variant/genotype calling. Each component would have specific physical, performance or other requirements that a developer would need to predefine. For example, the developer would need to consider the type of specimens that are needed as well as their volume and quality; or how the specimen needs to be collected, stored, shipped, or processed. Likewise, the read-length and coverage needed could dictate the type of platform used to generate sequence data. Developers would also need to consider limitations when designing a test, such as the availability of the necessary sample type or the ability of software analytical tools to detect the specific variants of interest. By considering each critical factor in the test development process and their impact on the overall design of the test, developers should be able to consistently generate high-quality NGS-based tests.

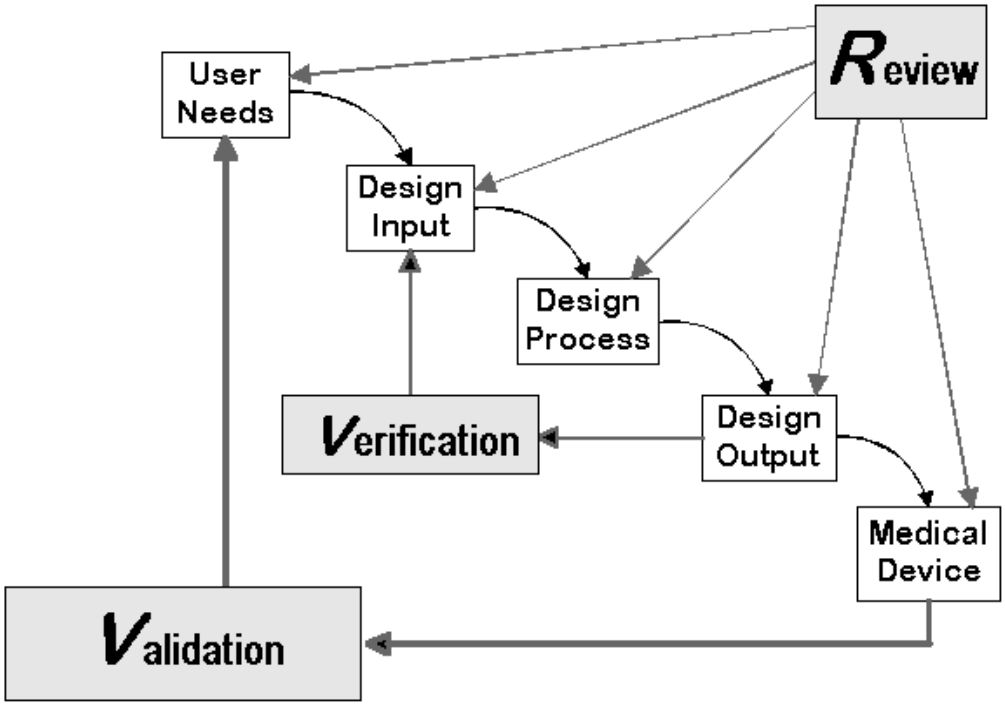

Figure 1. Conceptual depiction of design concept. From FDA guidance document "Design Control Guidance For Medical Device Manufacturers"
http://www.fda.gov/RegulatoryInformation/Guidances/ucm070627.htm

When considering specific design elements and critical factors, examples of the types of questions that a test developer may need to consider under such an approach would be:

- What is the purpose of testing? Is it for disease diagnosis of symptomatic individuals, treatment selection, non-symptomatic carrier testing, prenatal testing, or other uses?
- Who will be tested and what are the population features? If specific disorders are targeted by the test, how will you estimate the prevalence of the disorder in the population of interest?
- What is the required turnaround time (i.e., the amount of time between sample collection and the reporting of test results) of the test?
- What specimen types will work/are necessary? What volume (minimum, maximum) and quality are necessary? How do specimens need to be collected, stored, shipped, and processed based on other specimen requirements?
- What sample preparation is appropriate and necessary for the test? What DNA extraction method can be/needs to be used (if applicable)? What DNA quality/quantity is needed to obtain the expected accuracy? What is the sequencing instrument DNA input and what methodology needs to be used to accurately quantify input DNA?
- What type of sequencing reaction preparation, target enrichment will be used (e.g. amplification, capture type), if applicable? Are there interfering substances (including matrix effects) that might reduce the ability to amplify or sequence? What assessment will be performed to check for contamination? What methodology will be used to assess library yield and quality? How will DNA fragmentation and size be assessed?
- What sequencing instrument will be used and what limitations does this have for other factors (e.g., coverage, multiplexing)? What are the expected platform-specific artifacts? What are the

acceptance criteria for each run? (e.g., depth and uniformity of coverage, quality scores, duplicate reads, cluster density)? How many samples can or need to be multiplexed? What is the effect of multiplexing on coverage? What controls will be included in each sequencing run? How will barcode composition be performed to avoid barcode collision? What read lengths are needed? What type of sequencing will be used (e.g., single-end/pair-end/mate-pair sequencing)? What are the expected distances between the read pairs?

- How will coverage be assessed? Is coverage greater for some genomic regions than others? Is this identified? What minimum and average coverage depth is needed? What is the coverage of known medically interpretable gene regions, and how are those defined? What percentage of on-target coverage is needed?
- Which software tools will be used for mapping, alignment, and variant/genotype-calling? What kind of sequencing (e.g., targeted sequencing, WES, etc.) can the bioinformatics pipeline analyze? What types of variants can the software tools detect and report on? Is the software appropriate for the type of variants that need to be detected for the test results' intended clinical use? Will the bioinformatics tools be run locally or from a remote service (e.g., cloud-based)? Are there already existing software tools in place or will software tools be developed by the test developer? If existing, are these tools capable of detecting the variants of interest as is, or do they need to be customized for the test?
- What reference genome will be used? Is the selected reference genome adequate to detect the variants of interest? What percent of reads need to be correctly mapped? In what region(s)? What regions might pose a problem?

Under design concept standards, the developer documents conformity with the standards, which includes demonstration that the test performs in a way that meets both acceptance criteria and user requirements. User requirements are understood to include both analytical and clinical validity. Design concept standards may be developed by FDA, by third parties (and recognized by FDA), or may leverage existing standards.

Performance Standards

Performance standards would establish the metrics and performance criteria that are necessary for the developer to address and the test to meet, and could also include prescribed validation studies that a developer would be expected to carry out in establishing the analytical and clinical performance of the test. These standards could be developed by FDA, or by one or more third parties and recognized by FDA. These performance standards would not specify the actual design of the test (e.g., use of a specific platform, chemistry, or software); instead, the developer would be free to specify intended use, design, and other parameters, as long as the final test could achieve the defined performance metrics. As with design concept standards, appropriate validation studies would be conducted to determine conformity to performance standards, with results documented by the developer. It is likely that different sets of performance standards would need to be used for each individual device intended use. Different clinical indications for testing, such as whether the test is intended for disease diagnosis in symptomatic population, treatment selection, non-symptomatic carrier testing, prenatal testing in high risk or general population, pre-symptomatic risk assessment of developing a disease, prognostic use, etc. may require

different performance standards.

When considering defined performance expectations, examples of the types of parameters that could have defined performance characteristics may include accuracy, precision, limit of detection, interference/cross-reactivity, and certain process quality metrics. A goal of using performance standards would be to ensure that marketed tests meet minimum defined performance characteristics with clinically useful confidence intervals. Another goal is to enable to developer to understand and communicate the limitations of the test regarding certain performance characteristics.
Types of studies needed to establish test performance could also be prescribed in standards, such as those published by CLSI and other standards development organizations.

See the accompanying paper entitled "Additional Information to Facilitate Discussion on Analytical Standards Approaches" for a listing of existing standards, standards under development, and questions about practices that FDA has collected through analysis of literature, other documents, and current practice.

Questions

1. Would either a performance standard based approach or a design process based approach as described here be sufficient to ensure the development of high quality NGS tests? Would a hybrid approach or a completely different approach be more appropriate?
2. What elements are essential for a design concept standard for NGS-based tests? Are there elements that should not be a part of such standard?
3. Are there any additional procedures, methods, or test practices warranted when designing an NGS test that were not considered in description of a design concept approach in this document?
4. What elements are essential for a performance standard for NGS-based tests? Are there elements that should not be a part of such standard?
5. Are there any additional procedures, methods, or test practices warranted when designing an NGS test that were not considered in description of a performance standard approach in this document?
6. For performance standards, are there performance metrics that do not require minimum thresholds? If so, what is the best way to determine a range of acceptable values/thresholds for each metric?
7. Would separate performance standards be needed for tests with different intended uses?
8. What types of samples can be used in lieu of clinical specimens to develop NGS assay and determine performance characteristics? What should the expectation be for whether clinical samples need to be used or whether reference materials, reference sample panels, and other well characterized samples can be used? For example, how can we appropriately weigh disease and variant prevalence when deciding on what variants to evaluate and which type of samples to use? What are the most pressing reference panel needs?
9. Should there be a minimum number of specimens required for sufficient accuracy or should the sample numbers be determined by setting up acceptable 95% CI for PPA, NPA, or TPPV? What about precision, should there be minimum sample and replicate numbers analyzed? If defining acceptable point estimates, 95% CI, and other acceptance criteria – how can they be best defined and what should they be?

10. To what extent can performance characteristics determined using wild-type alleles be considered acceptable in lieu of variant alleles? To what extent can the performance using known non-pathogenic alterations be translated into performance for known pathogenic alterations? Would NGS test accuracy be adequately assessed if evaluated by variant class or genetic region instead of reported variants?
11. FDA has traditionally used well-established methods such as bidirectional sequencing as an acceptable comparator to establish performance of a new genetic test, which may not be feasible for all NGS tests. What comparators may be best suitable to evaluate NGS test accuracy?
12. Confirmatory testing, using Sanger sequencing or another appropriate orthogonal method, is currently routinely performed for germline NGS testing. What metrics, studies or acceptance criteria would be required to not have to perform confirmatory testing of all reported calls?
13. For developers specializing in providing bioinformatics data analysis, what information should be provided that is not covered here?
14. When the testing model relies on third party bioinformatics analysis, are there activities/documentation/other that the test developer should be required to address to ensure that the test results are accurate?
15. Are there additional issues FDA needs to consider for somatic testing? Are there additional issues FDA needs to consider for whole genome sequencing?
16. Can computational solutions to evaluate process quality and performance of NGS tests be developed?
17. What are the areas with the most pressing needs for analytical standards for test development and evaluation? What are the gaps in currently available reference materials and methodological standards and guidelines (e.g., missing, incomplete standards) and how should gaps in standard needs be prioritized?
18. What might be the most suitable and efficient models for developing new standards and are there groups already working to develop some of the needed standards? What groups should develop methodology standards to fill the identified gaps – FDA, standards organizations, other groups? Are there ways to accelerate the process?
19. What might be the best way to address modifications, software updates, etc.? How can we incorporate flexibility for modifications yet define appropriate specifics of tests and pipelines? What may be the best way for standards and evaluation approaches to be iterated upon and updated as knowledge and technology evolves?

Analytical Performance Parameters

Accuracy metrics. Positive percent agreement (PPA), negative percent agreement (NPA), “technical” positive predictive value (TPPV).

Precision **-** presented as a mean and the associated 95% CI, broken down by variant types tested, the number of replicates for each, and what conditions were tested. Precision includes repeatability (within-run variability), and reproducibility (taking into account all major sources of variability, such as run, reagent lot, instrument, operator, site, etc., as applicable).

Process Quality metrics such as:

- Pre-sequencing quality metrics such as genomic DNA concentration, volume, and quality, library yield, fragmentation, size range;
- Sequencing / base calls potential sources of error detected by examining run, cluster density, base call quality scores, number of reads, percent pass, cluster passing filter rate;
- Mapping or assembly (post sequencing) metrics - percent of reads correctly mapped / mapping quality scores, uniformity of coverage, average coverage depth, minimum read coverage depth;
- Variant calling metrics - variant call quality score, percent heterozygous calls, variant allele frequency (e.g., expected call frequency thresholds should be defined for homozygous and heterozygous calls, vs mosaic or contamination), number of reads with the variant reported, systematic errors, portion and ratios of base substitutions (transition/transversion), percent of novel variants, allelic read percentages, potential areas that contain large numbers of false positives or false negatives, concordance rates with reference variant/sequence, percentage of claimed region covered / percent completeness (i.e., percent of test with sufficient coverage above minimum threshold, vs percent of test with insufficient coverage).