

Analytic Validation: NGS Tumor Genomic Profiling - Julia Elvin

Female Speaker:

So, today we're going to be discussing the role of analytic validation in Next-Generation sequencing tumor genomic profiling. And let me congratulate all of you who have dialed in for such an exciting-appearing topic, and to stress how complicated this topic actually is. And the fact that we're going to spend some time together beginning to scratch the surface of the intricacies is just very applaudable, and I hope that after this conversation it will inspire you to ask questions and continue to dig.

So today, what we're going to cover is an initial overview of the technical challenges of doing tumor genomic profiling and why this is particularly germane to the specifics and extent of the analytic validation that each individual web performs. And we're going to go into depth on exactly what we mean by analytic validation and who in the regulatory community currently is evaluating the quality and extent of each lab's analytic validation, and then some clinical implications, and then some questions at the very end that hopefully will guide your discussions in the future. So let's frame up this discussion with a clinical case, which I think really emphasizes what we all hope to see within our lifetime, the promise of personalized cancer care.

Here are two patients who have diffusely advanced metastatic melanoma, and who have failed all forms of conventional therapy. The concept in personalized cancer care is that the tumor cells depend on abnormal signaling and growth for survival, and this is related to genomic changes, changes at the DNA level, which have driven these cells to become different from the normal and well-behaving cells in their body. So the first step is to identify the genes of interest that have been mutated and that are potentially producing the protein target. This is where the diagnostic tests comes in and is so important in discriminating next steps for patients.

Step two is then to treat with small molecules that inhibit abnormal pathways, or hit the Achilles' heel of the tumor while preventing [unintelligible]. I'm sorry, there's a lot of noise on the line -- treat with small molecules that inhibit the abnormal pathways within these cells, and spare the patient the vast majority of side effects. So in this rubric, this patient had an accurate test that identified a genomic alteration in each case, which allows the prescription of the medication that's specifically targeted to their tumor and resulted in a dramatic change in the course of their illness. So in this case, the rubric of diagnostic test predicting a specific therapy works well. But this is not as simple as it might appear, because we have not just one drug for one genomic alteration, but we have over 150 intracellular targets that we know have compounds in the pharmaceutical industry in development that will be hitting the clinics either in clinical trials or for use in actual treatment of patient cases in the next five to 10 years. So getting diagnostic pairing correct has never been more important.

So we've now established that there are molecular technologies that are moving into the clinic to predict responsiveness to drugs, and that patients' physicians are going to rely on these results for clinical decision-making. And with this complexity of potential therapies, labs and healthcare as a whole have moved towards multiplex technology to more broadly assess the genomic drivers, and this has added a level of complexity that we've never experienced before to effectively divide patients into the most relevant groups for whatever clinical interventions, which may actually mean withholding a particular therapy when a particular driver is not present.

In order for this rubric to be successful, we have to separate two very important concepts; the concept of technical variability that happens at the level of the assay. This must be minimized so that we can begin to understand the inherent biologic variability present in each patient's tumor. If these two sources of variability are not able to be separated, this rubric is not as effective at predicting what patients should go on a particular therapy. So let's take a look at an example of a Next-Generation sequencing assay workflow.

So, people talk about Next-Generation sequencing and sometimes think of all of these tests just because they employ this technology known as NGS as one uniform type of test, but that's not the case. So this describes only a single set that's in the middle of this sequencing -- this workflow diagram, illustrated and enumerated here as the Illumina high seq under the word "sequencing" in the middle. The assay itself is actually comprised of a number of upfront processing steps that starts actually at the moment that a specimen is procured in an intervention in an outside hospital, and doesn't end until an informative clinical report is issued.

What we have to capture here is there are multiple steps. It's a highly complicated workflow with multiple steps that each have to be validated and understood, and in particular, there has to be an appreciation for the pre-analytic variables that are outside the control of the NGS laboratory. These include the fixation of the specimen, the procedure that was used for collection, the age of the specimen, the storage condition if this is an archival piece of tissue, and this needs to be evaluated so that each individual assay understands the impact of each of these on the results that will be eventually delivered.

The second level or issue that we need to kind of address that is part of the complexity of a validation scheme is that there are not just one or two analytes that need to be evaluated. This is just a selected list of a number of genes that have been recognized, and that the presence or absence of a mutation directs a patient to a particular therapy or away from a particular therapy. The issue that compounds this list that we're adding to every day, and the scientific and medical literature is adding to every day, is the fact that any one of these genes listed there can be altered not just in one way by a point mutation that people typically think of when they invoke the word "mutation," but can be altered in four discrete and different ways.

So here, we have a cartoon which illustrates one mechanism, which is copy number alteration where the gene sequence itself is totally normal, but we just have, instead of two copies in the normal diploid cell, have more than two copies. This is the case that occurs in HER2-amplified breast cancer that we know is a target for herceptin therapy and other drugs. The second mechanism is base substitution. This is, again, the most commonly considered alteration, and this is where one area of a sequence, or one base in a sequence, is changed to another one, and this leads to a change in the coding and the amino acid of a sequence of the eventual protein.

The third category of alteration are insertion/deletion events in which small areas, particularly potentially regulatory areas, are either duplicated or deleted, or possibly insertions or deletions that are in non-multiples of three, which cause frame shifts in the coding sequence of the protein that eventually lead to early termination of the protein and a lack of a complete sequence being produced by the cell. And finally, rearrangement, where two pieces of the chromosome

interchange with one another, forming chimeric proteins that function in novel ways, and sometimes drive tumor cell growth. And this is the case for the EML4-ALK fusions in lung cancer that you, I'm sure, are aware.

The third complication to this entire process is that the testing needs to be performed on the clinically available specimens that patients are having collected in the course of routine care. And as we've moved towards smaller and smaller biopsies with minimally invasive procedures that have less recovery time and complications associated with the sampling, the tests need to accommodate for this lower input amount of material as well as the routine processing that occurs in a diagnostic pathology lab, such as formalin fixation and the effect that it actually has on the nucleic acid within the cell.

And finally, an analytic validation has to address and understand the fact that in routine care, sampling of the tumor is often dominated by the background normal cells from the patient, and that only a small fraction of any piece of tissue that is collected has tumor in it. And so what we see here represented from an experiment that we did in our laboratory was to look at the relationship between the purity of the tumor, i.e., the relative proportion of the extracted DNA that's coming from the tumor cells -- because what you have to understand -- in this type of an assay, all of the DNA from all of the cells that are present in a particular sample are being extracted together to form the input that's being sequenced, and the frequency or -- sorry, the sensitivity that different techniques have for detecting alterations is affected by the amount of tumor that goes in to begin with.

So, in this example, if there is almost 100 percent tumor, a heterozygous mutation like a dominant base substitution mutation, a KRAF [spelled phonetically] alteration, could be detected at -- it would have a mutant allele frequency of 50 percent in a 100 percent pure tumor, meaning half of the DNA coming from this tumor would have this alteration, and the sensitivity in a standard capillary sequencing assay -- this is the older type of Sanger sequencing -- would only be 93 percent. So the detection is less than 100 percent purity.

If we drop to a 40 percent pure tumor where the mutant allele frequency would be 20 percent or less, the old version, currently considered the gold standard, would only have about a 55 percent sensitivity for detecting these alterations. In reality, what we need to do is to be able to detect alterations at a very low tumor purity and mutant allele frequency, or at least to understand the performance characteristics of an assay to know when we have to caveat the results that would be negative to also include the concept that it could be a false negative because of the mutant allele frequency falling below the limited detection of the assay.

So, how are we going to address these challenges in oncology, Next-Generation sequencing, and genomic profiling, especially in light of the fact that there are numerous different tests being used by different laboratories? How do you assess the ways in which they are different from one another? Well, especially since they may have different genes that are being analyzed, they may have different amounts of each gene that's on the test being assessed. They have different approaches to enrichment of the particular genes that are on the assay -- PCR versus hybrid capture -- and each of these approaches has implications for the sensitivity and specificity of the test. Which instrument is being used? This is kind of the concept of which box do they have,

from which manufacturer. And then, which types of those four mutations can be detected, and in what clinical context?

On top of all of it, how do we know that one of these various approaches has value and can be trusted to provide accurate and reproducible results for a patient population? Well, here's a comment from the NIH-DOE task force on genetic testing which says, "The reality is that there is no assurance that every laboratory performing genetic tests for clinical purposes meets high standards." With this framing, what are you to do? Well, a lab should start, in order to understand the product that they are producing and the results that they'll be delivering, with an analytic validation. So what is this? This is a process by which you determine whether an assay is able to discriminate the presence or the absence of an event that it was designed to detect. These have two basic components, a measurement, or an assessment of accuracy versus precision.

Accuracy; you can envision this as darts being thrown at a dartboard. So accuracy is how often does the dart hit the bulls-eye, and how close is that. That is described by measures such as sensitivity, the ability to correctly identify patients who have a disease; specificity, the ability to correctly identify the patients who don't have the disease; and then, based upon the prevalence of a particular condition in a population, the positive and negative predictive values of the test, in this particular patient with a positive result or negative result, how likely does this reflect the actual status of the patient. Precision, on the other hand, is the concept, if you were throwing darts, how well do the darts cluster, even if they're nowhere near the goal line. So this is the measure of how much random variation there is in a test, and it's described by reproducibility and repeatability.

Why does analytic validation matter? Well, 70 percent, roughly, somebody has been heard to say, of medical decisions are based on the diagnostic test results of one kind or another, and these results stratify patients into subsets which get very different types of interventions or counseling. So the analytic validation helps assess the reliability of the data that's being given to the clinicians, which is feeding their medical decision-making. So, right now, who evaluates analytic validation? In general, there are a few organizations that provide this assessment and licensing, but in the area of Next-Generation sequencing, there is no single standard or guideline which regulates what is the gold standard for an analytic validation. So we're just going to be briefly go through these agencies to describe their role in this environment of regulation.

CLIA is the minimum bar that allows a laboratory to deliver tests which will prompt clinical decision-making, and they're charged with ensuring accurate and reliable test results. They inspect laboratories on an every-two-years basis, and will do a review of both the tests, which are FDA-cleared and -approved and being utilized in the laboratory, but also the laboratory-developed tests. So these are tests that the lab has either assembled from other or modified from other FDA-cleared or approved products, or has just generated entirely on their own to meet a clinical need. Within CLIA, there are no minimum thresholds that must be met specific to NGS testing.

The second regulatory body to consider is the College of American Pathologists. This is a credentialing agency that laboratories can voluntarily subscribe to for inspection and

accreditation. The inspections are also performed on an every-two-year basis. And mainly these checklists try to assess the quality management and quality control of lab testing, personnel, and lab safety. They have recently added some molecular pathology-specific checklists with sections that address Next-Generation sequencing, validation, and the ongoing Q.A. and Q.C., but the recommendations are fairly broad and open to quite a bit of interpretation.

The next regulatory body is New York State Clinical Laboratory Evaluation Program. This is a license which is required by every laboratory that wants to perform testing on New York State residents. And they have elevated the bar for licensing of molecular tests and Next-Gen Sequencing to acquire a New York State license, and it's currently considered one of the most rigorous certifications that a test can go through outside of the FDA. Recently, the Palmetto MoIDX program has also, in this lack of sort of clear regulatory guidance, established some components that they will be evaluating in their technical assessments specifically around the analytic validation of NGS-based tests in order to qualify them as covered tests. These components include sensitivity, specificity, and precision, and when it has gone through this technical assessment and has been deemed to be covered, they'll be listed on the MoIDX website.

And finally, FDA. So FDA typically is informed of tests and the performance characteristics before a test goes to market if the test is going to be FDA-approved or -cleared. However, they have been practicing enforcement discretions with regard to the laboratory-developed tests for many years, and in reality, most genetic and genomic tests are not FDA-approved products, but are lab-developed tests. And because of this, and the implications for the pairing with FDA-regulated drugs, this is a keen area of interest, and has led to a draft LDT guidance that was issued in October of 2014 and a diagnostic test workshop that included a number of thought leaders, just about a year ago, to advise the FDA on how they should proceed.

So, the next two slides just shows some comparisons between the New York State MoIDX and [unintelligible] guidance around things that you would want to consider in a validation. New York State is the only one that specifies how many clinical specimens they want to see results on for licensure, and this is just 50 specimens. The analytic sensitivity and specificity requirement guide MoIDX includes assessment of a limit of detection to be established for the minimum amount of DNA that is input into a test. New York State also wants an assessment of the lowest mutant allele frequency that can be detected by a particular test, and this is going to become more and more pertinent to oncology specimens as we continue to have low tumor purity samples where the risk of false negatives is very high, and as we proceed into an era where patients will have exposure to targeted therapy and will subsequently develop sub-[unintelligible] alterations that are resistant mutations, which will indicate that the patient should cease receiving a particular drug.

Some other things that are specifically addressed in these guidance's include precision, the stability of the sample and reagent, reference integrals, and some quality control issues that need to be in place. And then the differences between these guidance's are listed here, and only New York State and CAP [spelled phonetically] have established key performance metrics for the entire process, from that beginning stage of extraction through data analysis. And as we already touched on, the lower limit of detection has been called out specifically as important by MoIDX. And finally and sort of surprisingly, the only group to address a positive or sensitivity control is

the New York State guidance. This is really important when we think of other lab tests, and the fact that we would be routinely doing a test without including a positive control for the assay seems pretty improbable if we were thinking about a chemistry, or a CBC, or some other blood test. So let's look at an example of an NGS validation for a complex Next-Generation cancer genomic profiling assay.

Here's an example, and this is the test that we run at Foundation Medicine where we're assaying 315 genes. The claims that are being made around this particular test include that the full exon coding sequences for the entire 315 genes are going to be assayed, and there are statements about the validated accuracy coverage and amounts of input tissue. So, how did we establish this? Well, we go back to this picture here of the Next-Generation sequencing assay workflow to say that in the absence of a regulatory environment that prescribed what to do, we had the opportunity and resources and the personnel here with the background to set up an extraordinarily rigorous validation that allowed us to understand our test performance characteristics in a way that is very important to the quality of the data that we are delivering. In brief, DNA and/or RNA is extracted from a block or a slide of formalin-fixed, paraffin-embedded tissue. This step and the pre-analytic variables that happen in the collection lab were evaluated for their impact on downstream processes and extensively optimized, as I'll show you in a future slide.

The next step is the DNA that is extracted from the sample is made into what we call a library. This is the genomic DNA that represents a mixture of all of the chromosomes, all of the DNA content that is present in both the normal cells and the tumor cells in that initial block of tissue. Now, how do we focus in on those 315 genes out of the thousands of genes that are possible in this library? And this approach is through a technique called hybrid capture. This is one of several sequence enrichment approaches that can be utilized, and again, this step was optimized and validated. Once this wet chemistry part of the assay is performed, it's loaded on what we all think of as the box, the Illumina high seq, and then the sequence that comes out the other end has to go through computational biology processes to call out the different mutations in any one or all of those genes that are on the assay. And each of these analytic pipeline algorithms also has to be validated for their accuracy in being able to pair the output from the sequencer to real events in a particular sample, and then this needs to be matched to a reference sequence. So even if we detect a difference in sequence, it needs to be matched to a database that describes the normal human variation versus those that are seen in tumor cells, and this leads eventually to a clinical report being issued.

So the next few slides are going to show you some data, not because I am going to spend very much time at all -- in fact, I'm only going to slide through these very quickly -- but to understand the complexity of the type of evaluation that needs to be performed in this type of evaluation. So here is the first slide, which is the impact of DNA extraction before and after optimization. And what this scatter plot shows is this is a 100 percent pure breast cancer tumor where no optimization has been performed. And if you don't know what you're looking at, it just looks like a bunch of dots, but when you compare it to what 100 percent optimized sample preparation shows, you can see that now we see discrete bands within the chromosomes extracted from this tumor, and this allows us to reduce the amount of tumor in the sample from 100 percent down to

what's more clinically realistic, 20 percent, and still pick out the loss of an important gene that's seen in lobular breast cancer, known as CDH1.

As I mentioned, we have to validate a variety of different things in this testing world. We have to be able to call out any alteration type -- substitutions, deletions, insertions, amplifications, and homozygous deletions as well as fusions -- at any position in the 315 genes, which is over a million bases of individual coding regions, and be able to detect it at any mutant allele frequency from one to 100 percent. So you can imagine the complexity of designing a positive control that would allow us to push the assay to let us evaluate all three of these parameters. And so there isn't a patient sample, there isn't a reference sample, that can be purchased, or wasn't at the time, that was complex enough to evaluate all of these parameters simultaneously, so what we did is we created a pool from a variety of cell lines where the DNA mutations were very well-known, and this allowed us to model schematic mutations. The beauty of this [unintelligible] approach is that there are tumor cells and matched normal cells from the same patients, so that by combining these in different ratios you can understand the performance of the assay down to very low or high mutant allele frequency.

But then these next few slides just understand the numerous, numerous experiments that we did to show that we could detect mutant allele frequency of less than five percent of the total DNA across a large number of different genes covered in the [inaudible]. And we also looked at the performance over various amounts of [unintelligible] coverage to understand where we needed to put a quality control cut-off, so that if our test wasn't performing on a particular day up to the specifications, we would repeat the test rather than releasing a potentially incorrect result. We repeated these over time and looked at the correlation between the measured mutant allele frequency and what was expected based on the proportion of the normal and the abnormal cells that were added to the mixture, and we saw a linear relationship.

We also did these same sorts of experiments for cell lines that had known insertions and deletion events to confirm we could detect these at different mutant allele frequencies, and we repeated this for copy number alterations. So, cell lines with mixtures of different homozygous deletions or amplifications of particular genes were also challenged in this way and repeated so that we understood that when we had a 20 percent tumor fraction, and a gene was amplified at eight copies or more, we had a sensitivity of detection of 93 percent, and this one up to 100 percent, if our tumor fraction was more than 30 percent of the total DNA we were extracting. This allows us to confidently give results on specimens that are staged at a tumor content of 30 percent, and to qualify results for patients who have a tumor fraction of less than 30 percent if we find reasons to believe that it's possible that a copy number call may have been missed in a particular sample.

We also tested our platform against other tests that were available on the market, such as [unintelligible], with a large variety of FFT [spelled phonetically] samples, and looked at the concordance between the cross of both of these channels. There was 97 percent overlap here, or overlap of 97 of the mutations, with a few more being culled [unintelligible]. And when we looked at the additional mutations, the ones that were in this area detected by NGS and not by [unintelligible] were the ones at lower than real frequency. So, likely our true calls that were below the lower limits of detection in the [unintelligible] platform.

We also tested the second fish [spelled phonetically] in [unintelligible] with excellent concordance, and ran multiple experiments to look at the reproducibility between the sequencing results from the same specimen in inter- and intra-batch comparisons, and we did this reproducibility over time, so months and months and months. So these are 79 and 71 replicates of two different tumor samples where we knew what the alterations were that we were looking for, and every time we culled the exact same alteration, and we culled them almost at the same mutant allele frequency, which is the little bit of zigzagging at the line that you see here. But in every case, all three alterations were detected, and this resulted in our ability to describe our analytic validation results based on sensitivity and positive predictive values across a range of mutant allele frequencies, i.e., the tumor content present in a specimen, for all categories of all alterations.

And we didn't just submit this to CLIA or CAP [spelled phonetically] or MoIDX, but submitted it to a group that had no stake in verifying these results other than scientific interest. So this was the submission and publication of the analytic validity to Nature Biotechnology. And the kind of scrutiny that the scientists have on these editorial review boards is much higher than the scrutiny that is performed at a regulatory level, and on top of it, they require you to submit all of the data so that they can go through it in a fine tooth comb and make sure that you've drawn the correct conclusions. So if you go to this publication, this also has extensive supplementary data that includes the raw sequencing and mutant allele calls here so that anybody can draw their own conclusions about the validity of the test.

So what are the implications to patient care? Here are three examples from our experience comparing in an ongoing quality assurance process internally for a subset of lung cancer cases. We know that the NCCN and a variety of other guideline-issuing agencies recommend both EGFR and ALK testing to be performed on patient samples to directly care, so we were curious to see if the samples where we had to identify an EGFR exon 19 solution which is known to activate this gene -- how many of these tests, or how many of these specimens were previously tested, and what were the results. Did they agree with what we had seen?

So we looked at a variety of cases; we had 250 of these where the pathology reports were available, and review them for the presence or absence of information, presence or absence of previous testing results, and this was available for 71 cases. We identified that 12 cases had prior negative testing results, which represents 17 percent false negative rates. And you might say, "How do you know that these were true positives, and not some false positives that were detected by the assay?" So the clinical information and treatment that followed for one of these patients supports these being true results. Here's a patient who benefited from empiric [unintelligible] despite the fact that she had been given a negative EGFR testing result that fell into this category of a positive result by the NGSFA in a previous negative result.

When we look at the less common alterations that are just outside of the classic range, 83 percent of these patients were missed by [unintelligible] methods. And, again, here's an example of a patient that responded to EGFR-targeted therapy. We repeated this evaluation, looking at the cases that we identified as being ALK-positive, and found that, similarly, about 32 percent of the cases we identified as being ALK-rearranged had been previously called negative by fish testing. Most importantly, of these patients who were then subsequently treated with crizotinib, 70

percent responded to therapy, which is the same response rate seen in patients where the [unintelligible] ALK results are positive. So, again, these are true biologic positives that were fish methodology negative.

And finally, here's an example of patients that were evaluated by a very well-known laboratory that has had testing for a variety of markers, all of ones that are in NCCPN guidelines, by a combination of modalities including hot spot testing and multiplex sizing assays, for EGFR, HER2, KRAF, BCRAF, and two other genes, as well as fish assays to identify ALK RAF1 and [unintelligible] rearrangement. These patients were negative for all of these markers by the prior testing results. When these samples were run on the NGS-based profiling assay, a quarter of them had alterations that were within the genes recommended for testing by the NCCN guidelines. So a quarter of these [unintelligible]-patients, the best in class standard of care testing at the time had alterations that had not been recognized. An additional 40 percent had alterations that allowed them to enroll in a clinical trial for a targeted therapy agent that was available at their treatment institution.

So, in summary, when you're thinking about the validation, you need to remember that the quality of the lab validation and their understanding of their performance characteristics very much impact patient care. So some key questions you might consider asking of a lab that's presenting you with the possibility of a test is, number one, does the lab either have a peer-reviewed, published analytic validation, or have they successfully completed the MoIDX tech [unintelligible]? If not, would the lab provide you with the lab data from their validation for review? Is the lab New York State-approved? Were the validation specimens that were utilized representative of actual patient samples? Meaning, are they complex enough and reflect the low tumor purity of samples they're likely to encounter in clinical testing samples? Did they validate all types of all alterations or variations --

Female Speaker:

Is someone sure if they're -- if we're going to be able to get the handouts?

Female Speaker:

-- that would be represented in clinical testing? Were the sizes of the validation set large enough, and were the statistics appropriate to ensure narrow confidence in [unintelligible]? Was the entire process, from extraction all the way through reporting, validated and to a degree that ensures reproducibility and robustness? If a comparator method was used, what was it, and it should be available? And finally, does this assay validation include intra-assay and intra-assay precision studies between different operators over multiple days?

[end of transcript]