FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

ONCOLOGIC DRUGS ADVISORY COMMITTEE

Tuesday, July 24, 2012

9:00 a.m. to 3:00 p.m.

FDA White Oak Campus

Building 31, The Great Room

White Oak Conference Center

Silver Spring, Maryland

Meeting Roster

DESIGNATED FEDERAL OFFICER (Non-Voting)

Caleb Briggs, Pharm.D.

Division of Advisory Committee and Consultant Management

Office of Executive Programs

Center for Drug Evaluation and Research

ONCOLOGIC DRUGS ADVISORY COMMITTEE MEMBERS (Voting)

Deborah K. Armstrong, M.D.

Associate Professor of Oncology

The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

The Johns Hopkins University School of Medicine

Baltimore, Maryland

Frank Balis, M.D.

The Louis and Amelia Canuso Family Endowed Chair for Clinical Research in Oncology

The Children's Hospital of Philadelphia

University of Pennsylvania School of Medicine

Philadelphia, Pennsylvania

Brent Logan, Ph.D.
Professor of Biostatistics
Division of Biostatistics
Medical College of Wisconsin
Milwaukee, Wisconsin

Mikkael Sekeres, M.D., M.S. ***(Chairperson)***
Associate Professor of Medicine
Staff, Cleveland Clinic Taussig Cancer Institute
Department of Hematologic Oncology and
Blood Disorders
Cleveland, Ohio

Antoinette J. Wozniak, M.D., F.A.C.P.
Professor, Department of Oncology
Wayne State University School of Medicine, and
Karmanos Cancer Institute
Detroit, Michigan

Jane Zones, Ph.D. (*Consumer Representative)*

Medical Sociologist (retired)

Breast Cancer Action

National Women's Health Network

San Francisco, California

INDUSTRY REPRESENTATIVE TO THE ONCOLOGIC DRUGS ADVISORY COMMITTEE (Non-Voting)

Howard Fingert, M.D., F.A.C.P.

(Industry Representative)

Senior Medical Director, Clinical Intelligence

Millennium, the Takeda Oncology Company

Cambridge, Massachusetts

TEMPORARY MEMBERS (Voting)

Aman U. Buzdar, M.D.

Vice President Clinical Research and Interim

Professor of Medicine

M.D. Anderson Cancer Center

Dept. of Breast Medical Oncology

Houston, Texas

Peter Choyke, M.D.

Chief, Molecular Imaging Program

Senior Investigator

National Cancer Institute

National Institutes of Health

Bethesda, Maryland

Ralph D'Agostino, Ph.D.

Chair, Mathematics and Statistics Department

Boston University

Boston, Massachusetts

S. Gail Eckhardt, M.D.

Professor and Division Head, Medical Oncology

Senior Associate Director, Translational and

Collaborative Research

University of Colorado Cancer Center

Aurora, Colorado

Tito Fojo, M.D., Ph.D.

Program Director, Medical Oncology

National Cancer Institute

Bethesda, Maryland

David Harrington, Ph.D.

Professor of Biostatistics

Department of Biostatistics

Harvard School of Public Health

Boston, Massachusetts

James Liebmann, M.D.

Assistant Professor of Medicine

Department of Medicine

University of Massachusetts

Worcester, Massachusetts

Musa Mayer

(Patient Representative)

New York, New York

Michael Menefee, M.D.
Assistant Professor
Division of Hematology and Oncology
Mayo Clinic
Jacksonville, Florida

Lalitha Shankar, M.D., Ph.D.
Chief, Clinical Trials Branch
Cancer Imaging Program
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

David Steensma, M.D.
Associate Professor of Medicine
Harvard Medical School
Boston, Massachusetts

Wyndham Wilson, M.D., Ph.D.

Chief, Lymphoma Therapeutics Section

Metabolism Branch

Center for Cancer Research

National Cancer Institute

National Institutes of Health

Rockville, Maryland

GUEST SPEAKERS (Non-Voting, Presenting Only)

Cindy Dinella, R.Ph., Pharm.D.

President and Managing Partner

Advyzom, LLC

Berkeley Heights, New Jersey

Daniel Sullivan, M.D.

Professor and Vice Chair for Research

Department of Radiology

Duke University Medical Center

Durham, North Carolina

SPEAKER (Non-Voting, Presenting Only)

Lori Dodd, Ph.D.

Mathematical Statistician

Division of Clinical Research, Biostatistics

Research Branch

National Institute of Allergy and

Infectious Diseases

National Institutes of Health

Bethesda, Maryland

FDA PARTICIPANTS (Non-Voting)

Richard Pazdur, M.D.

Director

Office of Hematology & Oncology Products (OHOP)

Office of New Drugs (OND), CDER, FDA

Anthony J. Murgo, M.D.

Associate Director for Regulatory Science

OHOP, OND, CDER, FDA

Rajeshwari Sridhara, Ph.D.

Director

Division of Biostatistics V (DBV)

Office of Biostatistics (OB)

Office of Translational Science (OTS)

CDER, FDA

Jenny (Jing) Zhang, Ph.D.

Statistical Reviewer

DBV, OB, OTS, CDER, FDA

C O N T E N T S

C O N T E N T S (continued)

P R O C E E D I N G S

(8:00 a.m.)

Call to Order

Introduction of Committee

DR. SEKERES: Good morning, everybody. I think it's the appointed hour, so we'll get started. I'm Mikkael Sekeres from Cleveland Clinic. I'm a medical oncologist. I'd like to go around the room and have each person introduce their self and provide your affiliation. This is an experienced panel, so I don't have to give too much coaching about the microphones. Just remember to press the "mic on" button before you talk.

We'll start over on my right side.

DR. FINGERT: Good morning. I'm Howard Fingert. I'm a medical oncologist/hematologist. And I'm from Millennium, the Takeda Oncology Company. And I'm the industry representative to ODAC.

DR. CHOYKE: Pete Choyke. I'm a radiologist at the National Cancer Institute.

DR. ECKHARDT: Gail Eckhardt, medical oncologist, University of Colorado.

DR. WILSON: Wyndham Wilson, medical oncologist, NCI.

DR. STEENSMA: David Steensma, oncologist at the Dana-Farber Cancer Institute in Boston.

DR. MENEFEE: Michael Menefee, medical oncologist, the Mayo Clinic, Florida.

DR. FOJO: Tito Fojo, medical oncologist, medical oncology branch, NCI.

DR. LIEBMANN: James Liebmann, medical oncologist, University of Massachusetts.

DR. BUZDAR: Aman Buzdar from MD Anderson, medical oncologist.

DR. BALIS: Frank Balis, pediatric oncology, The Children's Hospital of Philadelphia.

DR. BRIGGS: Caleb Briggs, designated federal officer, ODAC.

DR. ARMSTRONG: Deborah Armstrong, medical oncologist, Johns Hopkins.

DR. WOZNIAK: Toni Wozniak. I'm a medical oncologist at the Karmanos Cancer Institute in

Detroit.

DR. LOGAN: Brent Logan, biostatistician, Medical College of Wisconsin.

DR. D'AGOSTINO: Ralph D'Agostino, statistician from Boston University.

DR. ZONES: Jane Zones. I'm a medical sociologist and the consumer rep. And I'm affiliated with Breast Cancer Action and the National Women's Health Network.

MS. MAYER: Musa Mayer. I'm a breast cancer advocate. I'm the patient rep for this meeting.

DR. ZHANG: Jenny Zhang, statistical reviewer, FDA.

DR. SRIDHARA: Raji Sridhara, division director, biometrics, FDA.

DR. MURGO: Anthony Murgo, oncologist at the FDA.

DR. PAZDUR: Richard Pazdur, office director.

DR. SEKERES: Great. Thank you, everybody.

We have a little bit of an unusual situation in that we have David Harrington, who's a professor

of biostatistics, in the department of biostatistics at the Harvard School of Public Health in Massachusetts, who is here by phone and represented by an empty chair over there. We're having some technical issues right now with his audio, so we're getting those cleared up.

For topics such as those discussed at today's meeting, there are often a variety of opinions, some of which are quite strongly held. Our goal is that today's meeting will be a fair and open forum for discussion of these issues, and that individuals can express their views without interruption. Thus, as a gentle reminder, individuals will be allowed to speak into the record only if recognized by the chair. We look forward to a productive meeting.

In the spirit of the Federal Advisory Committee Act and the Government in the Sunshine Act, we ask that the advisory committee members take care that their conversations about the topic at hand take place in the open forum of the meeting. We are aware that members of the media

are often anxious to speak with the FDA about these proceedings. However, FDA will refrain from discussing the details of this meeting with the media until its conclusion.

I would like to remind everyone present to please silence your cell phones and other electronic devices, if you have not already done so. The committee is reminded to please refrain from discussing the meeting topic during breaks or lunch. Also, as a reminder, on today's schedule, there is no scheduled break during the morning. We should be heading straight toward the lunch session. If I get the sense that attention is waning or if there's a need biologic breaks, then I will intervene, and we'll have a short break.

Now, a conflict of interest statement will be read by Caleb Briggs, the designated federal officer for the Oncologic Drugs Advisory Committee.

Conflict of Interest Statement

DR. BRIGGS: Thank you. I'd first like to recognize the press officer, Chris Kelly. If you're here, could you please stand?

Thank you, Chris.

The Food and Drug Administration, FDA, is convening today's meeting of the Oncologic Drugs Advisory Committee under the authority of the Federal Advisory Committee Act, FACA, of 1972. With the exception of the industry representative, all members and temporary voting members of the committee are special government employees, SGEs, or regular federal employees from other agencies and are subject to federal conflict of interest laws and regulations.

The following information on the status of this committee's compliance with federal ethics and conflict of interest laws covered by, but not limited to, those found at 18 U.S.C., Section 208 and Section 712 of the Federal Food, Drug and Cosmetic Act, FD&C Act, is being provided to participants in today's meeting and to the public.

FDA has determined that members and temporary voting members of this committee are in compliance with federal ethics and conflict of interest laws. Under 18 USC Section 208, Congress

has authorized FDA to grant waivers to special government employees and regular federal employees who have potential financial conflicts when it is determined that the agency's need for a particular individual's services outweighs his or her potential financial conflict of interest. Under Section 712 of the FD&C Act, Congress has authorized FDA to grant waivers to special government employees and regular federal employees with potential financial conflicts when necessary to afford the committee essential expertise.

Related to the discussion of today's meeting, members and temporary voting members of this committee have been screened for potential financial conflicts of interest of their own, as well as those imputed to them, including those of their spouses or minor children and, for purposes of 18 USC Section 208, their employers. These interests may include investments, consulting, expert witness testimony, contracts, grants, CRADAs, teaching, speaking, writing, patents and royalties, and primary employment.

Today's agenda involves the evaluation of radiographic review in randomized clinical trials using progression-free survival, PFS, as a primary endpoint in non-hematologic malignancies. They will consider the merits of an independent audit of investigator progression assessment in a prespecified subgroup of patients instead of an independent review of all progression assessments.

The expectation is that an independent audit would streamline the conduct of clinical trials, as well as avoid missing data when no additional protocol specified progression assessments are mandated. Hematologic malignancies are excluded from this discussion because other issues, e.g., blood counts, lymph node exams, and other biomarkers, influence the assessment of PFS.

This is a particular matters meeting during which general issues will be discussed. Based on the agenda for today's meeting and all financial interests reported by the committee members and temporary voting members, no conflict of interest waivers have been issued in connection with this

session. To ensure transparency, we encourage all standing committee members and temporary voting members to disclose any public statements that they have made concerning the product at issue.

With respect to FDA's invited industry representative, we would like to disclose that Dr.Howard Fingert is participating in this meeting as a nonvoting industry representative, acting on behalf of regulated industry. Dr. Fingert's role at this meeting is to represent industry in general and not any particular company. Dr. Fingert is employed by Millennium Pharmaceuticals.

With regard to FDA's guest speakers, the agency has determined that the information to be provided by these speakers is essential. The following interests are being made public to allow the audience to objectively evaluate any presentation and/or comments made by the speakers.

Dr. Daniel Sullivan has acknowledged that he is a scientific advisor for Covidien Pharmaceutcials on their R&D Advisory Board. He receives honorarium for one meeting per year.

Dr. Cindy Dinella has acknowledged that she is a consultant for Aragon Pharmaceuticals, Delcath Systems, Hoffman-LaRoche, and Synta Pharmaceuticals as part of her consulting company, Advyzom. As guest speakers, Drs. Sullivan and Dinella will not participate in committee deliberations, nor will they vote.

We would like to remind members and temporary voting members that if the discussions involve any other products or firms not already on the agenda, for which an FDA participant has a personal or imputed financial interest, the participants need to exclude themselves from such involvement, and their exclusion will be noted for the record. FDA encourages all other participants to advise the committee of any financial relationships that they may have with the firm at issue.

Thank you.

DR. SEKERES: Okay. I'd like to ask Dr. Shankar to introduce yourself, for the record.

DR. SHANKAR: Good morning. I'm Lalitha

Shankar. I'm the chief for clinical trials in the cancer imaging program at NCI.

DR. SEKERES: Great. Thank you. And I believe Dr. Harrington is on line now.

Dr. Harrington, could you introduce yourself for the record?

DR. HARRINGTON: Thank you very much. This is Dave Harrington from Dana-Farber Cancer Institute. I'm a biostatistician.

DR. SEKERES: Great. Thank you.

We'll now have brief opening remarks from the FDA.

Opening Remarks - Rajeshwari Sridhara

DR. SRIDHARA: Good morning, Mr. Chairperson, members of the ODAC committee, ladies and gentlemen. I am Rajeshwari Sridhara, division director of Division of Biometrics V, in the Office of Biostatistics at CDER. Today's meeting is unique in that we are not asking the committee's advice on merits of a specific drug or biologic product, but the purpose of this meeting is to have a wide-ranging discussion and advice on

how best to assess and mitigate potential bias in the determination of disease progression in non-hematologic malignancies. We will not be discussing whether progression-free survival, or PFS, is an appropriate efficacy endpoint or what magnitude of PFS benefit would lead to a marketing approval.

Progression-free survival, or PFS, is defined as the time from randomization to either disease progression or death, whichever occurs first, where an event is either progression or death, and PFS time is censored in patients who are alive with documented progression at the time of analysis. Disease progression determination can be made using both clinical and radiographic evaluation.

In clinical trials that are used to establish efficacy, we have mainly considered radiographic progression in the determination of PFS. When PFS is the primary efficacy endpoint of a clinical trial, FDA has generally required review of radiographs by an independent radiologic review

committee, or IRC, under the assumption that local evaluation or investigator assessment, INV, could potentially be biased. Please note that IRC and blinded central review committee, or BICR, are used interchangeably, and similarly, investigator and local site evaluator, or LE, are used interchangeably.

An extreme example was discussed in April 2011 ODAC, where examining the same radiographs, the investigator and IRC had diametrically opposed recommendations regarding the reason to terminate the trial. The investigator recommended trial termination based on the conclusion that improved efficacy had been established, whereas the IRC recommended study termination based on futility. Thus, the role of IRC is to mitigate potential evaluation bias by investigators. However, this approach may lead to a greater than 30 percent disagreement at patient level between investigator and independent reviewer assessments and are among independent reviewers themselves.

Because treatment is generally changed after

investigator determines progression, resulting in no further protocol-specified progression assessments, this practice results in missing data and informed censoring for IRC-determined PFS analyses. These disagreements have been attributed to a variety of reasons, including monitoring different target lesions.

In order to further examine the role of IRC, in October 2009, FDA, in collaboration with groups representing the Drug Information Association, the Pharmaceutical Research and Manufacturers of America, or PhRMA, and the National Cancer Institute, conducted a workshop on PFS to examine the discrepancies in PFS determinations by investigators and IRCs.

A meta-analysis of 27 trials conducted by the PhRMA working group indicated that while discrepancies in determining the progression dates can be observed, on an average, in 50 percent of patients, the relative treatment effect measured by hazard ratio between the experimental treatment and control are similar when assessed by either

investigator or IRC.

An inherent measurement error exists in the reading of radiographic scans, and disagreements between readers at the patient level are commonly observed. However, regulatory considerations are based on the relative treatment effect at the population level. These results and the results of FDA analysis to be presented here question the utility of complete-case IRC assessment and whether a random sample-based audit by the IRC can evaluate any potential investigator bias.

At the 2009 workshop, the NCI working group presented a plausible approach to such auditing. In order to confirm the meta-analysis results conducted by PhRMA, we at the FDA conducted meta-analysis of 28 phase 3 trials in 9 non-hematologic malignant indications submitted to the agency between 2005 to the present time. These trials included both investigator and IRC assessments of progressions. Of the 28 trials, 7 were in metastatic breast cancer, 7 in renal cell carcinoma, 4 in metastatic colorectal cancer, and

10 other indications, including non-small cell lung cancer, pancreatic neuroendocrine tumors, soft tissue sarcoma, gastrointestinal stromal tumor, ovarian cancer, and carcinoid tumors.

As a result of several trials having multiple cohorts or multiple treatment arms, the number of analysis units, or randomized comparisons, was greater than the number of trials, i.e., 33 for PFS and 30 for objective response rate. These trials included a variety of design features, as shown in this table. There were 13 open-label and 15 double-blind trials; 20 had 1-to-1 randomization, 8 with 2-to-1 randomization. There were trials with active control as comparator. There were add-on trials, placebo or best supportive care controlled trials, and substitution trials. Trial sample sizes ranged from 175 to 1,725 patients. While these trials may not represent the universe of clinical trials that are conducted, these do represent the clinical trials that are submitted to FDA for regulatory consideration.

The results of these meta-analyses are shown in these figures. On the X axis is the hazard ratio for PFS as determined by investigator, and on the Y axis, we have the hazard ratio as determined by the IRC. The redline is the line of perfect correlation. Circles above the line suggest relative treatment effect by IRC to be smaller than that of investigator, whereas circles below the line suggest relative treatment effect by investigator to be smaller than IRC.

All the circles are close to the line, demonstrating that the two are highly correlated, with a correlation proficient of 0.95. The panel on the right differentiates the open-label and blinded studies, with the green circles showing the blinded trials. The correlation was similar in both groups.

This graph depicts the correlation between investigator and IRC-assessed hazard ratio by indication. The correlations ranged from 0.87 for metastatic breast cancer, depicted in green squares, to 0.99 for renal cell carcinoma, depicted

in red circles. The blue triangles are the metastatic colorectal cancer trials, and black diamonds denote all other indications. In general, the correlations were similar across indications.

We also examined the investigator bias in the evaluation of objective tumor response using the trials which had reported both investigator and IRC response assessments. Again, on the X axis we have investigator-determined treatment effect as measured by odds ratio, and on the Y axis, we have IRC-determined odds ratio. We observed that, in general, at individual levels, the investigator-determined response rates in each of the treatment arms were higher than those determined by IRC. However, the relative effect as measured by odds ratio was similar, and in most cases, the effect in fact measured by IRC was larger than that by the investigator, as depicted by many circles above the perfect correlation line. The panel on the left shows the comparison in all the trials included in the meta-analysis, and on the right, they're differentiated by whether the

trials were blinded or open label.

This graph depicts the correlation between investigator- and IRC-determined objective response rate by indication. The correlation coefficient ranges from 0.81 to 0.98. The green squares denote metastatic breast cancer; red circles, renal cell carcinoma; blue triangles, metastatic colorectal cancer; and black diamonds, all other indications. More variability was observed in objective response rate compared to progression-free survival, as seen by the scatter of points around the perfect correlation line.

From these FDA conducted analyses, we conclude that there is high degree of association between investigator- and IRC-determined PFS effect. Assuming heterogeneity between the trials, when we evaluated using linear regression model weighted by trial size, the ratio of the hazard ratios of IRC versus investigator was 1.03. That is a 3 percent difference in the hazard ratios. While the objective response rate results were supportive, IRC is needed to mitigate potential

investigator overestimation of response. Given these results, a complete-case IRC for PFS may not be necessary and alternating methods such as a random sample-based IRC audit to evaluate bias must be explored.

In today's ODAC deliberations, the FDA requests the committee to consider the following points in their discussions. In order to have a fruitful discussion, we have invited speakers to present on potential audit strategies, measurement error standardized process and procedures in radiological measurement of disease progression, and logistical and feasibility considerations in conducting an audit.

Currently, two methods have been proposed for this type of audit. Dr. Dodd from NIH will follow me with her presentation of the audit methodology proposed by the NCI group. Her group proposes to evaluate the consistency of treatment effect as measured by hazard ratio between the IRC audited assessments and the investigator assessments.

Dr. Amit will represent industry's PFS working group and present their proposed audit methodology. Their method proposes to evaluate the differential discrepancy rates of investigator versus independent review committee between the treatment and the control arms. This will be followed by FDA's presentation of the evaluation of these two methods by Dr. Jenny Zhang.

While we are presenting two audit methodologies today, we expect in the future there may be other approaches for consideration. We recognize measurement errors, reader variability concerns, et cetera, exist in assessing radiographic progression, and we have invited Dr. Sullivan from Duke University to present on issues with the process and procedures of radiologic scans.

We acknowledge that current day clinical trials are conducted worldwide and bring in complexities, and we have requested Dr. Dinella, president of Advyzom Consulting Group, to present industry regulatory perspectives regarding

logistics of conducting an audit. We also request that the committee not focus their discussions on whether PFS is an appropriate endpoint or the magnitude of PFS effect that could lead to regulatory approval.

Given that regulatory decisions are based on the relative treatment effect and the observed high degree of correlation between investigator and IRC-assessed PFS relative treatment effect, we request the ODAC committee to discuss the following questions.

Given the information provided on random sample-based audit strategies, the variability in radiographic measurement, and logistical considerations, please discuss whether the current practice of complete-case IRC review of all patients should be replaced by a random sample-based IRC audit. Second, please discuss situations where a random sample-based IRC audit may not be appropriate. Thank you.

DR. SEKERES: Great. Thank you so much.

I'd like to invite Dr. Dodd up to give her

presentation.

Speaker Presentation - Lori Dodd

DR. DODD: Hello. It's a pleasure to be here today to present research on the use of progression-free survival and blinded independent central review in oncology trials. Most of these thoughts were sparked by my attendance at the December 2007 ODAC that voted on matters concerning the approval of bevacizumab plus paclitaxel in first-line metastatic breast cancer. The endpoint was progression-free survival. And an issue that was brought up during this meeting, amongst many, was the difference between locally evaluated progression times and those based on centralized review.

Concern was expressed about the high disagreement rates between central review readers and local assessments. They were around 30 percent. But importantly, the hazard ratios were in close agreement. I had worked closely with radiologists at the National Cancer Institute for nearly six years, and rates of discordance of 30

percent or greater seemed in line with what I had seen elsewhere in radiographic reading in oncology. The meeting raised many important issues about the use of central review, and the research I will present today addresses some of them.

I would like to point out for those on the panel that there are a few wording changes in my slides. They're very minor. And I've been told that the version that will be posted online is the version that I'm presenting now. I would also like to give a word of encouragement to the less statistically inclined. A lot of this is fairly technical that you'll be hearing this morning, so please stay with us because we need your input greatly.

So I was asked to talk about my research on auditing of PFS with blinded independent central review. There are many arguments for and against the use of progression-free survival as a definitive endpoint in a phase 3 trial. We must acknowledge that the use of PFS is an area of active debate. In general, it is not a measure of

clinical benefit. It does not directly measure how a patient functions, feels, or behaves, nor has it been generally shown to be a surrogate endpoint for overall survival.

We could spend all day and on into happy hour -- and I don't think we'd be very happy at that point -- discussing whether PFS is the right endpoint, and we must continue actively debating these of PFS. But for the purposes of my talk and for the day, let's assume that we agree that PFS is an important primary endpoint for regulatory approval.

Let me pause here for a moment to emphasize that a trial with PFS as an endpoint requires considerable evidence about the magnitude of the effect size. We want to be confident that the PFS hazard ratio is considerably better than just any improvement. For example, we may want to demonstrate a minimum improvement of PFS of 1 month, say rather than accepting anything better or greater than zero months. Sample sizes should be determined to give with reasonable precision

upper bound of the hazard ratio. The reason why I bring this point up is I'll return to it in the discussion of the audit.

So what is blinded independent central review? Progression assessments are evaluated with radiographs either by a local evaluator or my an independent central reviewer. Note that I prefer the use of the term "local evaluator" rather than "investigator assessed" because oftentimes assessments are made by the local site radiologist rather than study investigators.

So let's go through a graphic representing two patients. In this example, patients are evaluated for progression every six weeks. Here we have a patient who's randomized to Treatment A. The local evaluations -- that didn't go up all the way. Anyway, the local evaluations are evaluated at week 6 and week 12, and they determine there's no progression and then at week 18, a progression is determined.

You can see under this graph the images are now sent to blinded independent central review. So

the treatment assignment of A is blinded to the central reviewer, and the radiologist at the central review site evaluates the images. And in this setting, the central review radiologist calls progression at the second time-point. And as we've already heard, radiologists do not always agree in their assessments, so this kind of pattern is not unexpected.

Now, let's consider a second patient. This patient is also randomized to Treatment A. This patient has progression determined by local evaluations at week 2 -- or at week 12, the second imaging endpoint, and no further images are taken. So these two images are sent to the central reviewer who is blinded to the treatment assignment. And if the central reviewer does not call progression for this patient, then we do not know the progression time for this patient as assessed by a central review. So the central review information about progression is lost for this patient.

Now, it's because it seems likely that the

local evaluation progression tells us something about the likely central-review-called progression time that this creates a problem in terms of a potential for informative censoring. So it's probable that if one more image had been taken, the central review would have called progression at the third time-point, and certainly by the fourth point, rather than at 30 weeks or beyond. And it's because the local evaluation progression tells us something about when a blinded independent central review progression might have occurred that this pattern of missingness creates a potential for bias in the estimates of the treatment effect.

So in a paper with others in the Journal of Clinical Oncology in 2008, we wrote about blinded independent central review, and we asked if this was an important design element or an unnecessary expense. We discussed the issue of potentially informative censoring, which I've already reviewed. This occurs because patients are managed by the local site, and patients are typically taken off study at the time of the locally evaluated

progressions. And because the local evaluation progression time contains information about the BICR progression time, then this creates a potential for informative censoring.

In this paper, we also discussed the problem of measurement variability in the progression assessments and point out that BICR does not eliminate the measurement variability problem. So we cited a 35 discrepancy rate between two blinded independent central review radiologists. So clearly it doesn't solve the evaluation problem with progression.

We then began to ask what was the impact of this measurement variability on the estimates of treatment effect. Well, the answer to this depends on whether the measurement variability is the same across treatment arms. So in a paper published in Clinical Trials, we evaluated the impact of measurement variability when it is the same across treatment arms, and we found that it was not a major concern. On the other hand, if measurement variability is greater in one arm, then there is a

concern about bias. So say, for example, evaluators tend to call progression early in the control arm, then this would make all of us question the results. And this is the motivation for blinded independent central review.

So to summarize, there are three concerns being discussed. The first is that knowledge of treatment assignment during local evaluations raises concerns about potential bias. The second two relate to blinded independent central review. There's a potential bias from informative censoring, and blinded independent central review does not eliminate the measurement variability. That said, there's not a lot of evidence in the literature about bias and the estimates of treatment effect, based on local evaluations. Furthermore, comparisons of hazard ratios between a local evaluation and BICR seem to generally agree, as demonstrated by two meta-analyses.

So given this, what are our options? Well, we could say let's go back to overall survival. It is not measured with error. We know when patients

die. But we've already agreed that we're in a setting where PFS is an appropriate endpoint. We could also say that because there's no evidence in bias in the local evaluations in the literature, the local evaluation hazard ratios don't differ much from blinded independent central review, then just use the local evaluations. In the case of double-blinded trials, this argument is the strongest. We could continue with complete-case blinded independent central review, but this is costly and time-consuming. So a compromise is to use blinded independent central review on a subset.

There are many approaches to going about an audit. What I want to discuss now is one approach that we've proposed. We propose an audit whose purpose is to demonstrate that the blinded independent central review hazard ratio is of at least some prespecified minimum size. In other words, the audit must prove the effect size is of a certain magnitude.

Before leading you through the algorithm for an audit, I want to pause to come back to this

concept of the clinical irrelevance factor or the CIF. I've already mentioned the concept of powering a trial to demonstrate something other than any improvement. This concept translates to what I term the clinical irrelevance factor or the CIF. The CIF describes a threshold above which a hazard ratio would be determined clinically insignificant.

Another way to think about this is to ask what is the smallest effect size that is acceptable or clinically meaningful? This is not a statistical judgment, rather it is a clinical assessment of how much of an improvement in PFS is needed for a meaningful result? When specifying, consideration should be given to a minimally relevant improvement in PFS. This could be determined in terms of the minimum improvement in the median progression-free survival time, and then translating that into the effect on a hazard ratio scale.

For example, let's say we want to demonstrate the experimental treatment has at least

a 1-month improvement in the median PFS. If the control median is 9 months, then this corresponds to a null hypothesis of a hazard ratio of greater than .9. I just want to remind folks that a hazard ratio of 1 means that the treatments are the same in terms of their efficacy. So anything less than 1 we're using to indicate improvement in the experimental arm.

So the audit we're proposing is a retrospective audit. We assume that all images are collected and archived for all patients at all time-points they're imaged in the study. The exact timing of when the central review starts can vary. We do not need to wait until the study has ended, but this is the way we've implemented it and what I will present.

So the first thing we ask is if the locally evaluated hazard ratio is clinically meaningful and statistically significant. We perform a central review on a subset of patients if this is true. We compute the BICR hazard ratio, and in our paper we propose a more efficient estimator than simply

estimating the hazard ratio on a subset. So this means we can -- we don't have to use as many subjects to obtain the results as we would if we just used the subset alone. And we do this by incorporating information from both the central review and the local evaluation.

Then we perform a hypothesis test, and the hypothesis test is simply whether the hazard ratio satisfies our clinical irrelevance factor. And if it does, if we can reject a null hypothesis, then we stop the audit procedure. If we fail to reject a null hypothesis, then we would proceed to a full central review and test the null hypothesis again This is a two-stage procedure requiring up to two hypothesis tests. And in practice, we implement this using a Hochberg procedure to control for the type 1 error rate.

We also have a formula for determining the audit size. The audit size depends on many of the standard quantities we're used to assuming when we make sample size calculations. It depends on the number of progression-free survival events. If

there are fewer events, then we'll require a larger audit. It depends on the magnitude of the effect. So smaller effects will require larger audits. The third factor is the clinical irrelevance factor or the CIF. So ruling out a smaller effect will require larger audits.

The fourth component is something that we're not as familiar with in determining sample sizes. It's the correlation between the blinded independent central review hazard ratio and the local evaluation hazard ratio. If there's a perfect correlation between the central review and the local evaluation, that would imply we need no central review because the central review is giving us the same information as the local evaluation, and lower correlations would then require larger audits. And these are all features of the audit size formula we propose.

So we applied the audit method to five randomized controlled trials. We obtained data from Bristol-Myers Squibb, Genentech, GlaxoSmithKline and ECOG through data-sharing

agreements. The analysis was all conducted by me or under my supervision within NIH. Four of the studies were in metastatic breast cancer, and one was in colorectal cancer. The sample sizes ranged from about 740 to 209. The hazard ratios ranged from effect as large as .48 in the first study -- the paclitaxel plus bevacizumab study -- to .77, and we applied the audit method to each of these trials. Each of these trials had a full central review, so we were able to use computers to simulate what the audit process would be to evaluate the operating characteristics of the audit procedure.

So here's a graphic to describe what we did. So every study had a full central review. We took a sample, and we used the sample size formula for determining the audit size, and we took an audit sample of that size. Then from there, we estimate the hazard ratio, and we tested whether we could reject a null hypothesis that our hazard ratio was greater than the clinical irrelevance factor. If we rejected the null hypothesis, we would stop.

And if we didn't, we would proceed to a full central review and test the null hypothesis again. This whole process was repeated 10,000 times for each trial as an exercise to evaluate the performance characteristics of the audit procedure. In practice, the audit would be conducted one time.

So I'll present results for one of the five studies for the purposes of time. This was the paclitaxel plus bevacizumab trial. The hazard ratio was .48. And the sample size -- the total sample size was 722. The second row is assuming the null hypothesis we're trying to test is a clinical irrelevance factor of 1. This is a standard null hypothesis in clinical trials, just saying that there's any improvement. The proportion of complete case audits in this case was 4.2 percent. So we very rarely went to a complete case audit during our computer simulation procedures of this trial.

The mean audit size was 28 percent, which corresponded to, on average, having a central review of about 200 subjects. And then as a

feature of the design, we always continued to a complete case central review if we don't reject at the first look. And so, of course, all of the time we were able to reject the null hypothesis that the hazard ratio was greater than the clinical irrelevance factor.

When we make our clinical irrelevance factor more stringent and set that factor to .9, which would correspond to showing there was roughly a minimum of a 2 and a half month improvement in the median progression-free survival time, we proceed to an audit more frequently. So 16 percent of the time we went to a complete case audit, which makes sense because we're requiring more evidence. And the mean audit size in this case was 37 percent, so we're requiring larger audits. And then on average, that corresponded to about 270 subjects having their full set of images reviewed by central review. And also in this case, all the time we rejected the null hypothesis and concluded that we had demonstrated a reasonable magnitude of the effect for PFS with a central review.

So in summary, in conclusion, blinded independent central review does not resolve all problems related to progression-free survival assessments. The audit is a reasonable compromise between a complete-case central review and no central review. The audit that I've proposed focuses on the estimate of the treatment effect, which is the hazard ratio, and a BICR audit is an efficient means of evaluating the robustness of the treatment effect estimate. Another point is that larger treatment effects will tend to have smaller audit sizes than smaller effects, which is also evident in the paper that goes through the other four trials that we collected data from.

In addition, this discussion only applies to progression-free survival in the phase 3 setting. I don't want anybody to start thinking about applying this to the phase 2 setting where the issues are quite different. I feel that blinded independent central review might not be necessary in double-blinded trials.

Finally, I want to point out that the BICR

audit requirements may differ when the reasons for the blinded independent central review -- so, for example, when progression is very difficult to assess -- which I have heard, in the carcinoid testing, there's quite a bit of measurement variability in assessing progression in carcinoid -- then the motivation for doing central review differs. And, therefore, the audit procedure that I've presented and an audit in general may not make sense. Thank you.

DR. SEKERES: Very good. Thank you so much, Dr. Dodd.

We'd like to invite Dr. Amit up to give his presentation.

Industry PFS Working Group Presentation - Ohad Amit

DR. AMIT: Good morning and thank you for the opportunity to present on this very important topic of progression-free survival and central review. The work I'm going to present here, as culminated over the last four years, is part of the PFS independent review working group. You see all the key companies that contributed to this effort

listed on the slide here and names of the key contributors. Some additional acknowledgments are several other contributors to this effort that I wanted to acknowledge before proceeding into the talk.

I think many of the important points have already been made by Dr. Dodd and Dr. Sridhara, but just to reiterate what I believe are the key points that I wanted to present today, firstly, as I'll show you in a few slides, we believe as part of our work that the local evaluation and the blinded central review provide very comparable estimates of the treatment effect in the great majority of clinical trials, and they've looked at this in the meta-analysis.

That said, there are still rare situations of course where evaluation bias may be present. And in those situations a blinded central review is a mechanism for auditing both the quality and the reliability of the local evaluation. But given that these are more rare situations, based on the data we've looked at, it's desirable of course to

lower the significant resource burden associated with a central review by developing methods for detecting evaluation bias based on a sample of patients or an audit. And what I'm going to present to you today in our methodology is based on differential discordance, which we believe is an effective tool for assessing evaluation bias in a sample-based procedure.

So just a bit more introduction before I get into the data. As I mentioned, I'm presenting this on behalf of the independent review working group. Many of the members are here. This working group was formed in June 2008. One of the first things we did was undertake a meta-analysis to evaluate concordance in the estimates of the treatment effect between central review and the local evaluation, and this was done across many solid tumors. I'll show you that data in a second. We then used the results from that meta-analysis to motivate and develop methodology for a sample-based independent review. The key findings are published in an article in the European Journal of Cancer,

and most of these findings is what I'll present here today.

So moving right into the results of the meta-analysis, this was very similar to what Dr. Sridhara had shown you earlier. Our meta-analysis was based on 27 trials across multiple solid tumors. Predominantly, it was metastatic breast cancer, colorectal cancer, and renal cell carcinoma. What you can see on the Y axis is the hazard ratio by independent review. On the X axis, you see the hazard ratio by investigator. And you see the yellow bubbles represent blinded trials, while the white bubbles represent open-label trials. And the size of the bubbles is directly proportional to the size of the trial. Once again, you see a very high correlation there of .947 between the treatment effects, and you see most of the points lining up around the 45-degree reference line. And we fit a new intercept progression line, which is almost similar to that reference line.

We also looked via a mixed model at the ratio of the hazard ratios, with 1 representing

perfect agreement. And you can see the estimate there across the 27 trials is 1.02 with very tight confidence intervals.

So moving on now, before I give you the details of our procedure, I want to talk a little bit about discordance. There are various ways to look at discordance. Fundamentally, at the patient level, it represents a disagreement between a local evaluation and a central review regarding either the occurrence or the timing of progression, and from there, one can calculate a discordance rate. That's the rate at which disagreements occur on either the occurrence or timing of progression within a treatment arm. And then the differential discordance, as we've defined it, is simply the difference between treatment arms and the discordance rates. There are various different ways to measure discordance. We can define it multiple different ways. Some are going to be more useful than others in terms of their value in detecting evaluation bias.

So moving on to talk about the goal of

central review, we believe the goal of any independent review, whether one does it in an audit or whether one does it in a full case review, the goal is really to confirm the treatment effect. So I state this in the presence of highly concordant estimates of treatment effect that we've observed. And we've observed these highly concordant treatment effects in the presence of significant discordance at the individual patient level.

So what does this mean when you observe very concordant estimates of treatment effect and still see a lot of discordance? I think this has been mentioned earlier, and it's worth restating. I think discordance is primarily a consequence, then, of measurement error. But that said, it can also be induced by evaluation bias.

So what is critical here to us in designing a procedure? We want a procedure that can separate the evaluation bias from the measurement error. When talking about evaluation bias, I think it's important to kind of note the mechanism by which this occurs. So fundamentally, I think what

happens when you see evaluation bias is the investigator or the local evaluation is systematically calling progression earlier or later on one arm relative to the other; so at a much higher rate on one arm compared to the other. And what we've noted, and what I'll show you in a couple slides here, is that this leads to different discordance patterns or rates in the experimental and control arms, and that's what we call differential discordance.

So how do we define and evaluate discordance for the purpose of our audit or a sample-based methodology? We've defined a couple of metrics of discordance shown on the bottom there. I just want to note, for the early discrepancy rate in the publication, we actually had b+a3 in the numerator there. In theory a3 is not observed very often, but it does in practice get observed. So for completeness, it should read b+a3 on the numerator there.

What you see up top is a 2x2 table outlining results by BICR and local evaluation, with the

off-diagonal elements representing the discordance cases. Also, a2 and a3 in the top left represent an agreement on the occurrence of progression but a disagreement on the timing. And so we can use this table to define very simple measures of discordance that are defined here. The early discrepancy rate essentially looks at the rate, that the locally assessed PD is called earlier than the centrally reviewed PD. And then conversely, the late discrepancy rate or the LDR is looking for the proportion of disagreements, where the local evaluated PDs occur later than the centrally reviewed PDs. I don't want to focus too much on these formulas. Just note that when you calculate these sorts of things, obviously if you have evaluation bias, you would expect the rates for these two measures to differ between the arms.

So just looking at differential discordance, what happens when there's no bias and when there is bias? On the left-hand side there, you see simulated data from 10,000 trials where there was no bias. And you can see the reference line, the

horizontal reference line there, represents ratio of the hazard ratios with 1 meaning perfect agreement between the IRC and the local evaluation. And what you can see there is almost an equal scatter of points above and below that reference line of 1.

Similarly, you see a vertical reference line of zero representing no differential discordance or no difference in arms between discordance. And you can see in the situation where there's no bias, you can see as many points to the right of the line as to the left, and what you're seeing there is predominantly measurement error.

We have data available from 12 clinical trials, and we were able to superimpose that data on this plot to show that you see a fairly consistent pattern with the simulated data when you look at discordance rates versus HR ratios and real-world data. Now, on the right-hand side what you see is a situation where we've imposed biased into the simulation. And now you can see firstly that most of the points -- most of the trials are

shifted to the right. So you tend to see a lot more differential discordance when you impose bias into the simulation. And similarly, a lot of the points are shifted above the horizontal reference line of 1. So you can see the effect of imposing bias.

The last point I wanted to make about this slide is also you can obviously see a very strong relationship between differential discordance and the estimates of the treatment effect by central review and local evaluation. So as you get more and more differential discordance or difference between arms and discordance, you see much more divergent estimates of treatment effect for the hazard ratio between central review and local evaluation.

So moving on to talk about our procedure, our proposed procedure now, what is the goal of our procedure? I think it may be slightly different than what Dr. Dodd presented, but I think, again, as she noted, there are several potential goals that one might want to have for an audit. The goal

of ours is to increase the confidence and the integrity of the trial and the trial endpoints. We're not proposing or intending to re-estimate the treatment effect from our audit or sample-based procedures.

So what are the key concepts that are supporting the audit methodology? Firstly, I think what allows us to move to this methodology is the fact that local evaluation historically we believe is providing good and reliable estimates of the treatment effect, and, therefore, we want to reduce the burden of central review, but still retain a mechanism for detecting meaningful bias in the estimates of the treatment effect. And that meaningful bias in the estimates of the treatment effect, as I've shown you on the last slide, manifests differences between treatment arms and the discordance rates, as measured by the two metrics that we've proposed.

So, operationally, what does the methodology look like? Quite simply, at the time that enrollment is completed, one would identify a

random sample of subjects. At the time of the clinical cut-off for the final analysis, central review would be performed in a random sample that was identified. And then one would proceed to break the randomization code, perform the analysis, and estimate the local evaluation, the hazard ratio for the local evaluation. At that point, one would also estimate differential discordance and compare that to some prespecified threshold value.

Based on that comparison, we would either conclude that the local evaluation hazard ratio is reliable or we would conclude that there may be some evidence of evaluation bias. And then we would move to a complete-case review or we would estimate the hazard ratio by central review.

So before I talk about the operating or performance characteristics about our procedure, I just want to make a little note because I think the FDA will present some data, subsequently, that are based on our original publication, where we define sensitivity and specificity as shown on the slide. At a high level, sensitivity is simply the

probability that we're going to detect bias in the sample when bias is truly present. And conversely, the specificity of the procedure is a probability of declaring the local evaluation reliable, given that no bias is present. And in our original publication, we also had a fixed threshold value that was fixed regardless of the sample size.

We've modified these definitions a bit, and I'll show you that in a second. Primarily, what we've done is we've benchmarked our sensitivity and specificity relative to what is the current practice for detecting evaluation bias, which is a comparison of the hazard ratios, based on the full case review. And what we're essentially saying in the full case review is, essentially, a relative difference of about 25 percent in the hazard ratios between the local evaluation and the central review would lead you to conclude that there's some evidence potentially of evaluation bias, and that's how we benchmarked performance characteristics.

So sensitivity, as we've defined it, and as I'll show you in the subsequent slide from the

simulated data, it's essentially the proportion of the time that we detect evaluation bias in the audit, given you would have detected it had you done a full case review. And conversely, specificity is proportion of the time we conclude the local evaluation as reliable in the audit, given a similar conclusion would have been reached based on the hazard ratios from the full case review.

So just a note on the sample size and the threshold values for the audit, we chose our threshold values and our sample sizes based on fixing the sensitivity for detecting bias at 90 percent. So we want to fix the sensitivity, which is arguably what most people would not want to lose, which is our opportunity to detect bias when it's present. We wanted to fix that at 90 percent, and our threshold value then becomes sample-size dependent. And depending on the desired specificity, sample sizes of 100 to 200 subjects are needed, so the specificity will increase as your sample size increases.

So here are the operating characteristics of our procedure. This is based on 10,000 simulations. For the LDR, you see again -- for both the LDR and the EDR, given that we fixed our sensitivity, the sensitivity is coming out right at around 90 percent. On the right-hand side, what you see are the threshold values, and you can see that they're increasing with sample size. So essentially as you increase the sample size, you're setting a higher bar; because you have more confidence, you're setting a bar to move to a full case review. And as you set that bar higher, you can see your specificity is going to increase as well. What I will also note from this slide is it appears that the LDR in terms of specificity performs a little bit better than the EDR.

So before I conclude, just a few notes on when we think a sample-based approach or an audit should be done. When trials are truly blinded, we don't believe that a central review is necessary at all. But when one has open-label trials or when complete blinding is not possible, then a

sample-based procedure or an audit is appropriate. There are still going to be situations, we believe, where 100 percent BICR is going to be desirable. Trials where sample size is smaller, I think the audit in that situation may not be feasible. There may not be a lot of logistical savings and one might just want to proceed to do a full central review if a central review is warranted. And there are going to be situations where one wants to increase the confidence in the local evaluation; for example, in tumors where RECIST criteria may be more difficult to apply.

So in summary, I think, firstly and most importantly, what we've seen is that the local evaluation is very consistently providing a reliable estimate of treatment effect, but there are still situations where bias may be present. And in those situations, we believe differential discordance is a useful tool for detecting evaluation bias. And it can be used to design audits with a manageable size and good operating characteristics.

Some next steps for our working group, we certainly want to apply our procedure a bit more retrospectively in some existing clinical trials, and we have plans to do that. And obviously there's a key step of regulatory acceptance which hopefully is part of the discussion that we are having here today. And with that, I'll close and say thank you for your attention.

DR. SEKERES: Great. Thank you very much.

I'd now like to invite Dr. Zhang on behalf of the FDA for her presentation.

FDA Presentation - Jenny Zhang

DR. ZHANG: Good morning. My name is Jenny Zhang, a statistical reviewer in the Division of Biometrics V, CDER FDA. I would also like to acknowledge my team members, Drs. Huanyu Chen, Lijun Zhang, and Raji Sridhara.

This is the outline of my presentation. I will give a brief summary of Dr. Sridhara's presentation as background and motivation, then go into the details of FDA's evaluation of the two previously presented proposed audit methods by

Dr. Dodd and Dr. Amit. Two cases studies will also be presented, including one study with definitive evaluation bias presence, and I will conclude with a summary.

As shown by Dr. Sridhara, FDA's meta-analysis of 28 prospective, randomized phase 3 registration trials in solid tumors corroborated the high degree of association between investigator and IRC PFS treatment effects purported in recent publications. This finding suggests that complete-case IRCs may not be necessary in many oncology trials and motivates the exploration of alternative methods for bias evaluation, specifically, audit methods.

The idea between the audit strategy is to increase our confidence in the investigator result of PFS by conducting an IRC review in a random sample of patients. The main savings of such a strategy lies in the situation where there is no actual bias in the investigator result and only a partial IRC audit is needed to confirm that fact. Other potential benefits include a reduction in

trial complexity, a reduction in cost and burden to investigators, the avoidance of some missing data issues, and mitigation of informative censoring, a main concern with IRC analyses.

Two currently available proposed audit methods that FDA has evaluated are those just previously presented and will be referred to herein as the NCI method and the PhRMA method. A brief summary of the NCI method is given here, where the goal of the audit is to provide assurance about the investigator PFS treatment effect estimate. Thus, an IRC audit should only be considered when the investigator hazard ratio indicates a clinically meaningful and statistically significant effect in favor of the experimental arm.

As mentioned in Dr. Dodd's presentation, a more efficient estimator of the IRC hazard ratio is proposed. A formula to estimate the audit size is also provided, which depends on factors including the effect size and what they call the clinical irrelevance factor or CIF. The CIF is a threshold value; for example, a hazard ratio equal to 1 used

in the proposed two-stage testing procedure. The upper bound of the confidence interval of the IRC hazard ratio estimate is compared to the CIF to determine whether consistency of the PFS treatment effect has been verified. Since all trials had a complete-case IRC conducted, random sample audits are performed 10,000 times for each trial to assess the performance of the NCI method.

The PhRMA method is summarized here. The basis of their method is to use differential discordance as a measure to detect evaluation bias. From this 2x2 table, two measures are defined. The early discrepancy rate or EDR is the frequency that the investigator declares progression, or PD, earlier than the IRC. And the late discrepancy rate or LDR is the frequency that the investigator declares progression later than IRC.

The differential discordance for each measure is the difference between the rate on the experimental arm and that on the control arm. The idea is that a differential discordance beyond a certain threshold is suggested of bias being

present in the investigator assessment.

In the PhRMA group's original publication, threshold values ranging from .075 to .1 and IRC sizes of 100 to 160 patients were recommended through simulation studies. As you've just heard, the PhRMA or PFS working group has since conducted more simulations, and as a result has modified its recommendations with respect to threshold values and audit sizes. However, since those new results were not available to FDA at the time of our evaluation, the results presented here will follow the publication recommendations.

With respect to the interpretation of the differential discordance measures, a negative differential discordance for the early discrepancy rate, or EDR, and/or a positive differential discordance for the LDR, or late discrepancy rate, are indicative of bias in the investigator results, in favor of the experimental arm. A negative differential discordance for EDR means a higher rate of investigator progressions being called earlier than IRC on the control arm, and a positive

differential discordance for LDR means a higher rate of investigator progressions being called later than IRC on the experimental arm.

The performance characteristics of these proposed audit methods need further evaluations and real clinical trial data to determine whether the audit strategy is a feasible alternative. We evaluated the two audit methods in 27 prospectively conducted, randomized phase 3 registration trials in solid tumors across 9 indications, as listed on this slide. Note that one metastatic breast cancer trial was excluded from these analyses due to aspects of the data not being conducive for analysis by those methods.

The table below summarizes the measures FDA used in its evaluation of the two methods. Recall for the NCI method, the sample audits are conducted 10,000 times for each trial to assess its performance so we can obtain summary measures of the mean audit size, the percentage of full audits, and the percentage of positive audits, where consistency of the PFS treatment effect is

concluded. Note that these replicate audits are conducted only for performance evaluation purposes and are not necessary when actually using the audit method.

For the PhRMA method, our evaluations calculated the differential discordance for both the early and late discrepancy rates, and we fixed the audit size to 160 patients. Recall that the recommended range of audit sizes from their publications was 100 to 160 patients. Analyses using other audit sizes were also performed by FDA and showed similar results.

This table summarizes the level of investigator and IRC discordance between treatment arms across the 27 trials, divided into disagreements on censoring status and the timing of progression within a 7-day window. We see that the discordance rates are very similar between arms and around 20 percent for both categories.

This plot assesses the NCI method by looking into the relationship between the mean audit size for each trial on the Y axis and the upper bound of

the 95 percent confidence interval of the investigator hazard ratio estimate on the X axis. The cluster of circles at mean audit size of 100 percent are those trials for which full IRC audits were needed in all 10,000 replicates. Those trials all had upper 95 percent confidence interval bounds of the investigator hazard ratios greater than .9. This means that, as expected, trials with borderline or non-significant investigator results would need full IRC audits.

For all other trials, mean audit size decreases with the upper bound. This means that trials with larger, more significant investigator results would obtain the most savings. In terms of meeting a much smaller audit size, most are below 50 percent.

This figure shows that the previously described general relationship between mean audit size and upper confidence interval bound holds across indications. These two plots assess the PhRMA method. By looking into the relationship between the hazard ratio ratio of IRC versus

investigator on the Y axis and the differential discordance for the early or late discrepancy rate on the X axis, obtained from a sample audit size of 160 patients, note that an HR ratio greater than 1 implies an overestimate of treatment effect by the investigator.

Recall that early discrepancy rate or EDR is the frequency that the investigator declares progression earlier than IRC. As explained previously, a negative differential discordance for EDR is suggestive of bias in the investigator result, favoring the experimental arm. In support of this rationale, we see that the differential discordance for EDR decreases as the HR ratio increases. This means that as more investigator progressions are being called earlier than IRC on the control arm, the difference in IRC and investigator hazard ratios also increases. The reverse relationship is true for the late discrepancy rate or LDR since LDR is the complement of EDR, that is, LDR is the frequency that the investigator declares progression later than IRC.

This figure shows that the previously described general relationship between the HR ratio of IRC versus the investigator and the differential discordance for EDR or LDR holds across indications.

This table summarizes the various measures from both methods by categorizing the trials with respect to their investigator hazard ratio estimate. Of the 12 trials, with a large observed investigator-assessed PFS treatment effect, that is, a hazard ratio of less than or equal to .5, the median mean audit size from the NCI method was 35 percent; and all trials resulted in positive audits, that is, consistency of the treatment effect was concluded.

The differential discordance for either EDR or LDR suggested bias in 5 of the 12 trials, or 42 percent, based on a random sample of 160 patients using a threshold of .075. For more moderate observed investigator treatment effects, the savings using the NCI method decreases to a median/mean audit size of 80 percent, whereas only

27 percent of these trials were recommended to go to a full audit by the PhRMA method.

One trial with definitive evaluation bias present was the carcinoid trial that was discussed by the ODAC in April of 2011. This was a phase 3 randomized 1 to 1, placebo-controlled study of everolimus for the treatment of patients with unresectable or metastatic carcinoid tumor. The primary endpoint was PFS by IRC. At their second interim analysis, an unprecedented discordance of the PFS treatment effect was observed between investigator and IRC. The investigator PFS result crossed the efficacy boundary while the IRC-PFS result crossed the futility boundary. Clearly, some bias was present in this trial.

The final results of this study are summarized in this table. The investigator PFS hazard ratio estimate was .78, while the IRC PFS hazard ratio was .93. The HR ratio, which is the ratio of IRC hazard ratio versus investigator hazard ratio, was 1.19.

It was of particular interest to FDA how the

two audit methods would perform for this study. The left table summarizes the discordance between arms seen in this study with respect to censoring status, progression time, and censoring time. We see some discrepancies between the two arms.

The right table presents performance results from the two audit methods. For the NCI method, 100 percent of the 10,000 replicates resulted in full audits, with zero percent being positive audits; that is, consistency of the treatment effect cannot be verified in any of the replicates. For the PhRMA method, however, neither the differential discordance for the early nor late discrepancy rate met the threshold to conclude that bias may be present, and a full audit was thus not recommended.

To illustrate the potential savings in audit size from the two methods, let's look at another case study. This was a phase 3, randomized, 1 to 1 placebo-controlled maintenance trial in 711 patients with soft tissue sarcoma. The investigator PFS hazard ratio estimate was .72, and

the IRC PFS hazard ratio estimate was .76. The HR ratio was, thus, 1.06.

The bottom table summarizes the discordance between arms. The right table presents performance results from the two audit methods. For the NCI method, only 14 percent of the 10,000 replicates resulted in full audits, with 100 percent being positive audits; that is, consistency of the treatment effect was verified in all the replicates. The mean audit size was 47 percent. For the PhRMA method, the fixed audit side of 160 patients was 23 percent of the total sample size. Using the threshold value of .1, bias is not present, and a full audit would not be recommended. Thus for such a study, at least a 50 percent savings in audit size could be obtained.

In summary, FDA's evaluation supports that a random sample IRC audit is a viable alternative to a complete-case IRC and may be a more efficient and cost-effective strategy to detect bias in the investigator results. The NCI method seems to perform well in those situations. In other words,

it seems able to distinguish between trials with and without bias present. However, the savings with respect to audit size varies from case to case. The PhRMA method is intuitively appealing, but needs further evaluation, particularly with respect to determination of the appropriate threshold value. This method may also suffer somewhat from a loss of important information due to dichotomization.

Selection of the actual audit strategy to implement within a trial may need to be determined on a case-by-case basis and difficult to generalize, however, this is an area of further research. These analyses have demonstrated that an IRC audit to assess potential bias in the investigator evaluation is a feasible approach.

I would like to conclude by thanking Dr. Lori Dodd for sharing her code for the NCI method, which greatly facilitated the timely completion of these analyses. Thank you.

DR. SEKERES: Very good. Thank you so much.

I'd like to invite Dr. Sullivan up to give

his presentation.

Guest Speaker Presentation - Daniel Sullivan

DR. SULLIVAN: Thank you. I've been asked to give a little bit of background about the issues that contribute to variability in tumor measurements and what might be done about this. And on my disclosure slide here, I note my work with the RSNA. It revolves around this issue. I coordinated activities called the Quantitative Imaging Biomarkers Alliance, which is focused on identifying the sources of variability and finding a means to mitigate or reduce them.

I'm going to focus just on CT today in the interest of time. Measurements can of course be made on MR and other modalities, but many of the issues are the same. I'm going to comment on three contributions to variability in tumor measurements: the image acquisition itself, the reader, the characteristics, and the measurement method.

On CT, there are a long list of technical factors which are known to influence lesion size, and, therefore, the anatomic response assessment.

And I won't go through all of these, but I'll just show you a couple of examples; and in addition, the patient himself or herself, depending on the phase of inspiration and whether or not the patient can suspend respiration. Because on modern CT scanners, the image can be obtained in a single breath hold, and whether the patient can or cannot do that makes a difference in blurring the margins.

These are some data from the literature of a couple of years ago showing representative scanners. The scanners m, n, o and p are scanners from the four major manufacturers that make CT scanners. They are all measuring the same reference nodule in this data. So the size of this nodule is known, and you can see the absolute percent errors here. They range from 7 up to almost 15 percent. And recently within the QIBA activities that I've just described, we have replicated these data on a wider range of scanners and find the same range of variability on modern scanners, up to plus or minus 15 percent.

These are two images -- these are the same

images showing the characteristics of different display. The image on your right, image B, has intentionally been displayed at a window level display to exaggerate the blooming of the water density elements in the image so that the margins of the tumor become obscured. On the image A, points a, b and c and d around the tumor are right on the edge, but on image b, they appear to be within the tumor. And as I said, this is exaggerated, but the radiologist can manipulate the image in such a way, inadvertently, that the margins of the tumor will change when making measurements.

Turning to the radiologists, there are a variety of characteristics. Whether it's a radiologist or an oncologist, or whoever, a technologist, whoever is making the measurements, one of the key issues is the level of skill or expertise that the observer has, also whether the reader has bias. And in particular, what comes into play is not necessarily bias about whether the radiologist knows the treatment status of the

patient on a day-to-day basis, but whether the reader has a bias to either under-call or over-call changes on the image because of the subjectivity. And I'll show you an example of that in a moment.

Measurement error, simple random discrepancies due to intra-variability and inter-reader variability. Lesion difficulty, whether the margins of the lesion are indistinct or obscured by other structures, and lesions with heterogeneous mixtures of density within them. Tracking different lesions, different target lesions, is another source, and overlooking the development of a new lesion if one is using the RECIST criteria.

As background for the next slide, where I'm going to show you radiologist variability, I want to start with this generic ROC curve. For those of you not familiar with how data or the performance of an observer are typically displayed in observing a signal, the receiver operating characteristic curve is a typical way to do this. And the Y axis is usually some measure of the true positive rate

or sensitivity, and the X axis is usually related to the false positive rate or some measure of specificity, usually 1 minus specificity in this kind of display.

This curve, the dotted curve, is connecting points that are called operating points of a particular radiologist who is categorizing a signal as to whether it is present or absent, to be used in a variety of settings. And the curve gives an indication of the particular skill of this radiologist. The curve that is higher up toward the upper left-hand corner indicates a radiologist, or a group of radiologists, with higher skill than curves that are lower. And the diagonal line represents the performance of someone who is just performing, according to random chance.

The points on this curve are referred to as the operating points for radiologists. For a radiologist emphasizing a specificity, then he or she would be operating toward the far left-side of the curve, down closer to the zero point. And a radiologist who is emphasizing sensitivity would be

operating at a point up toward the upper right-hand corner of the curve.

On this slide, these are operating points of 108 radiologists, and you'll have to superimpose in your mind the ROC curves that might correspond to them on this graph. These data are about 15 years old, but this graph is frequently referred to because it is considered to be a statistically valid sample of a group of radiologists performing the same study as is actually performed in the field. In a sense, those of you in this room would be familiar with a phase 4, postmarketing study for a drug. This would be analogous to a phase 4 study of a diagnostic procedure.

So these radiologists are operating at different operating points and have different skill levels. This is the line of random chance again. And as the radiologist is performing closer to the upper left-hand corner, he or she is operating at a higher skill level than other radiologists who are closer to the diagonal line. Radiologists operating at a different point on the curve, toward

the lower zero point or up toward the upper left-hand corner, that radiologist is displaying a difference according to value judgments. The variability in radiologists is some combination of these two components. It's very difficult to tease these apart, and it's also difficult to change these on a short-term basis.

This is an example of how this might play out in measuring tumors. The image on your left is pre-treatment, and the image on the left is post-treatment. One radiologist may outline the tumor for measurement with this yellow line. And this is an example of a radiologist operating at a high specificity level, so somewhere down on the left-hand side of an ROC curve. This radiologist wants to be sure that every pixel that he includes truly represents tumor. And he is not including pixels outside that might represent inflammation or scar, or something of a different level of certainty.

This is a line from another radiologist, and these are lines that were drawn actually by

radiologists. This is not a demonstration that I created. This is an example of a radiologist who would be operating at a high sensitivity level, somewhere near the upper right-hand corner on an ROC curve. This radiologist wants to be sure to include every possible pixel that might be tumor, erring on the side of including lots of false positive pixels, in a sense.

In one other example of differences, this is an example where reader 1 includes this area of this sliver of density heading towards the hilum, which may be scar or tumor, and another radiologist over on the far panel has outlined the tumor but excluded that, assuming that it perhaps represents a scar.

Thirdly, the measurement method is a source of variability, whether the radiologist is using a ruler, electronic calipers, automated techniques; the number of lesions chosen for measurement, measuring different lesions at different time-points; the choice of measurement -- unidimensional, bidimensional,

biometric -- and the human interaction, the radiologist's interaction with whatever device or methods, particularly software algorithms, that might be used.

At the most extreme, in large global trials, measurements might be done in a very coarse way with actually a physical ruler or a piece of paper, or relating them to a scale that is in the image shown there at the arrow. Hopefully this doesn't occur too much in clinical trials nowadays. More commonly, the radiologist would use software built into almost all CT scanners. Here, there are green crosses placed on each side of a lymph node, and the software in the CT scanner calculates the distance as 26.14 millimeters in this example. And virtually all CT scanners have some facility such as this, but it will differ from manufacturer to manufacturer.

A way to standardize this would be what's referred to as third-party tumor measurement software, which would operate on a different workstation. The scans would be transferred to the

workstation, and then the radiologist would make measurements in a similar way as I just showed you, but the software would be standardized for all scans from all manufacturers. Here, again the radiologist has made marks on this lymph node in this case. In the longer dimension, this would be under RECIST 1 rather than RECIST 1.1 with a maximum diameter listed here as 19.4 millimeters.

A more automated method is shown here, where the radiologist, instead of having to actually mark the margins of the tumor, puts a point, sometimes called a seed point, somewhere near the center of the tumor. And the software then assesses the characteristics of that pixel or voxel, determines all of the similar pixels or voxels that are similar to it and draws a margin around the tumor, in this case this magenta line, which would occur in multiple slices so that the software would then calculate diameters and a volume as well. And you may not be able to see, but there are two faint green lines crossing this tumor. The intersection of those two lines is at the third arrow that I

just put up there. And the software displays the diameters as 3.92 in one direction and 3.11 in the other, and also gives volume.

Although this is called automated, I've labeled this as automated, notice that the radiologist actually does have to be involved and start the algorithm, and has to then accept the result. So there are in fact, to my knowledge, no fully automated, FDA-approved, commercially available software programs that don't require an observer to be involved at some level. So there is some variability of that observer's interaction with the software.

I won't spend much time at all on this because lots of data has already been presented. There is a large body of data for various tumor types, body regions, modalities, acquisition parameters, and linear or volume measurements. And we have already heard today about the high discordance rates between two blinded, independent viewers in over 27 retrospective analyses. These are a few examples.

There are many published studies where the site and central reads disagree, and yet the treatment effect is not obscured. There are not so many published examples where they disagree, but one example has just been referred to by Dr. Zhang, the everolimus trial for carcinoid. And Lori Dodd mentioned in her presentation that carcinoid is a particularly difficult tumor to measure. And so it may be that there is not a one-size-fits-all method and that there may have to be some customization for tumors of certain sites.

Problems with trying to minimize these issues in standardized site reads, especially for large phase 3 trials; the difficulty in cost of training radiologists because training to mitigate the effects of skill or value judgments actually is difficult and takes time; difficulty in cost of auditing sites, especially in distant geographic regions; difficulty in cost of mandating a standardization and training in trials with more than 100 sites, each of which may only contribute a few patients.

Then I wanted to mention that there are inadequate software standards for recording segmentation and measurement results with images. I showed several examples where radiologists had put marks on the images on the CT scan or with third-party software, with automated software. And there are not good software standards to capture that information for future auditing and to maintain an audit trail. There are standards in the works, but they're not yet widely disseminated or readily available.

I'll quickly review the RECIST recommendations related to this. In non-randomized trials where response is a primary endpoint, confirmation of PR and CR is required to ensure that responses identified are not the result of measurement error, also to be able to compare this with historical data. However, in randomized trials phase 2 or 3, or studies where there is stable disease or progression of the primary endpoints, confirmation of response is not required since it will not add value to the interpretation

of results. However, elimination of the requirement for response confirmation may increase the importance of central review to protect against bias, particularly in studies which are not blinded.

So some things that can be done or considered to reduce discordance or variability, a single reader should evaluate all exams for a given patient. The images should be provided to the reader in the clinical sequence in which they were obtained. The reader should choose the same lesions on each study; should choose the right lesions, and by that I mean measurable and not difficult. Choose measurements that are robust. More automation of measurements will help to reduce variability.

There could be improvements in rigorously defining non-target progression. RECIST does not have very good definitions at present for non-target progression of lesions, and there could be better development of CAD algorithms to automate that decision about non-target progression; for

example, algorithms that look at the texture changes within an image. Improvements to detect new lesions so they don't get overlooked. Again, there are computer algorithms that can assist with that in a lot of settings. And better response criteria, moving away from the four categories of RECIST to continuous criteria might also help to reduce discordance.

In addition, implementing scanner calibration and QA programs at each clinical site is essential, and there are a variety that can be used. There are also existing accreditation programs that go beyond just scanner calibration and QA. In particular, there is the NCI Quantitative Imaging Excellence program. And I listed their categories below to note that it focuses on volumetric MR and volumetric CT. It does not address linear measurements. And this is a reflection of the fact that the imaging community believes that volume is a better measure and that we should be moving that way. And essentially, the imaging community does

not feel that we should be exerting more effort and resources on improving linear measurements. That's consistent with QIBA activities as well. In addition to the NCI program, the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network also has a site qualification program.

So I mentioned QIBA, and just in the last minute, I want to just explain what that is. Our mission is to improve the value and practicality of quantitative imaging biomarkers by reducing the variability across devices, patients, and time, and we issue two types of documents to do that. One is an image acquisition protocol, which is similar to what you're all familiar with, for an image acquisition protocol describes the process for creating medical images. And it could be changed as needed for different clinical trials for different reasons. But in addition, we go beyond that to a document we call a profile, which is a systems engineering document that describes a specific performance claim and how it can be achieved. It's a more rigorous document, and it

cannot be changed, or you will not be able to achieve the claim that it states.

So, for example, in a CT of volumetry, we have in the past two or three years been doing what we call ground work, data collection, which is essentially reproducing lots of the data that you've seen but in a more comprehensive and standardized way to assess intra- and inter-reader variability of nodules of known size to determine the minimum biological change using clinical scans, using readers -- and these are readers who are used by the imaging CROs -- and then assessing the variability across all scanner models and sites in a comprehensive way. And building on that, we are now looking at the differences amongst algorithms, which all purport to provide volumes from these data. And the next step, then, will be to correlate with clinical endpoints and outcomes, which we have not done yet, but that would give us the threshold for clinical utility.

The current claim in our CT volumetry profile states that "a measured volume change of

more than 30 percent for a tumor provides at least a 95 percent probability that there is a true volume change." The fact that that 30 percent sounds similar to the 30 percent in RECIST is just coincidence. They don't have the same implication of change in terms of actual tumor volume. This claim holds when the tumor is measurable; that is, the tumor margins are sufficiently conspicuous and geometrically simple enough to be recognized on all images, and the longest in-plane diameter of the tumor is 10 millimeters or greater.

The threshold for actual clinical significance is to be determined. There are some people who think that it is sufficient to say that anything greater than 30 percent represents progression, but that needs to be validated because there are yet no accepted response criteria for volume from many professional organizations.

In conclusion, important efforts have been made to standardize image acquisition across sites, devices, and time to minimize subjectivity and interpretation and to improve consistency of

radiologic endpoint assessment, but endpoint evaluation is still influenced by scan variability, by the individual reviewing the image, and the time-point at which he or she reviews it. Thank you.

DR. SEKERES: Great. Thank you so much.

For our final presentation of the morning, I'd like to invite Dr. Dinella up.

Guest Speaker Presentation - Cindy Dinella

DR. DINELLA: Good morning. My name is Cindy Dinella. I'm the president of Advyzom, a boutique regulatory consulting firm. Previously, I was at Hoffman-LaRoche for 20 years in Nutley, New Jersey. I was U.S. head of regulatory. I've had the privilege to work with FDA in the oncology division for the last 18 years. I have seen the evolution of oncology development, regulatory endpoints, and approval of new treatments to advance patient care. I want to thank Dr. Pazdur and the division for inviting me here today.

As far as background, a brief summary of why we're here today, in part due to where we've been

in our collective best thinking, our learnings over time, and today with an assessment based on clinical trial experience to date with regard to IRC as a PFS endpoint, independent radiologic review is implemented under the assumption that investigator assessments could potentially be biased. We do see, and for expected reasons, a discordance rate that can range from 15 to 30 percent, and as noted by FDA this morning, up to 50 percent. However, the important outcome and regulatory hurdle is clinically meaningful treatment effect for the overall study, where no systematic bias can be detected.

PhRMA and FDA analyses of trials over time have observed a high degree of correlation between IRC and investigator-determined PFS treatment effects without systematic bias introduced by investigator, and there's been a number of publications on this. Today, I'd like to give you a collective industry perspective in IRC, focusing on value, burden, and regulatory need, with particular focus on bias, increasing trial

complexity, and cost.

In order to do this, FDA had asked Advyzom to objectively and broadly as possible collect and present industry feedback in conducting IRCs. Myself and my partner Krishnan Viswanadhan reached out to individuals and had individual discussions with sponsors from small, medium and large companies, as well as key consultant experts and members of the PhRMA working group. The list is on slide. There were some who wanted to remain anonymous, but I wanted to thank everyone for participating over a very brief period of time to enrich this presentation.

The collective feedback was very consistent. We believe investigator assessment should be considered the primary endpoint in randomized PFS trials. If and when needed, an independent audit of random samples of scans, according to pre-set criteria, is an important next step and needed. Primary reasons for change are burden, cost, and value. Burden and cost is inherent to drug development but worth it when the value of

high-quality data and endpoints provide the best answer. The question today is whether full IRC has that substantial additional value or does it duplicate efforts?

IRC versus investigator assessments has a balance. This slide is representative of the balance between the advantages and challenges of IRC versus investigator. I'll review them in detail within the presentation, but the evolution of knowledge for IRC over time has provided us some insight to its own challenges. At the end of the day, our collective view is the following: IRC does not represent more the truth, but IRC and investigator are two ways that evaluate PFS.

Just a perspective on bias, past and present. IRC originally was implemented to reduce investigator bias but does not totally eliminate bias in itself. IRC has the potential for bias and variability, as the experts have presented this morning also. First point, investigator-determined progression leads to missing data in IRC reads and informative censoring. The investigator determines

progression independent of an IRC read. Once progression is determined, no further follow-up scans may be provided to IRC, and patient may be crossed over, go on to other treatments, or be lost to follow-up. Therefore, the common practice is to censor these patients at the time of last tumor assessment, which can lead to bias in results.

The second point is the selection of different lesions or missing a new lesion development can lead to discordance amongst IRC readers themselves as well as to the investigator. IRC readers may assess different target or indicator lesions, which would lead to an adjudication process who would pick one of their assessments. But also the investigators themselves may be following a different set of lesions. Additional discrepancies can occur if a PD is called by the investigators themselves for a small emerging lesion which the IRC did not detect but is called a responder.

The third point is the variability in training and inconsistent application of RECIST

criteria. This has been seen in IRC reads. IRC readers can be involved in multiple trials with multiple tumor types. The critical importance I think, as outlined by Dr. Sullivan, is really the application of the RECIST criteria and the understanding of the details of it within a trial itself. Training needs to be just in time. Overall, IRC was developed for a good purpose of the potential investigator bias, but in itself has practical issues associated with it.

So the perspective today on bias is FDA analyses, PhRMA analyses, and the published literature by experts indicate no systematic bias has been introduced by investigators. To be fair, and based on some anecdotal discussions, I want to represent, the CROs or vendors who conduct some of the IRCs are concerned that investigator will be present and possibly worth the cost and burden due to the rigorous methodology employed. However, the truth may be that both IRC and investigator assessments have potential bias, albeit different. What's most important, and as noted by FDA and

other experts this morning, is the overall study outcome for PFS to meet regulatory hurdles as measured by investigator IRC has been comparable and demonstrated over the evaluation of a number of clinical trials.

In summary, perhaps we need to recognize that, over time, the education, the training, the rigorous regulatory standard that was set in place by FDA, and the overall results highlighting the importance of objective radiologic evaluation, has led to a successful performance of all stakeholders over time, and that includes the investigators. The original intent of IRC was to reduce the potential investigator bias, however, I guess the question is whether there was ever evidence from the beginning or today that indicates there has been systematic investigator bias in randomized PFS trials.

The totality of clinical data is in the hands of the treating physician, including reasons for withdrawals that include toxicity, so both clinical data and radiologic evidence is of

critical importance. Perhaps systematic bias is not an issue as deemed by the experts today. There's an advantage to use the investigator assessment in the totality of clinical data for which the assessments are based on.

So we're proposing to hopefully discuss and conclude today -- through our learnings, through the clinical trial results, through the expert analyses presented this morning -- that the investigator assessment could be rigorous and unbiased to allow for investigator assessment to be the primary regulatory endpoint with some controls in place, such as the audit.

Just a reminder, and I think as noted by Dr. Dodd, the real regulatory outcome is still for survival. Progression is an intermediate endpoint. PFS is an important intermediate endpoint, as it allows for seeking answers in a smaller patient population and allows for shorter trial duration due to the data progression occurring earlier than death. PFS is not confounded by the effects of subsequent therapies and, if treatment effect is in

fact substantial, can be clinically meaningful. But overall, survival is the ultimate endpoint, and investigators have more information about the patient in clinical progressions, which makes the investigators' call more relevant for survival than the central reviewers'.

The next topic, if all things were considered equal and today you can see that investigator-assessment bias has not been detected, then we have to go into the reasons for complexity of cost and burden. Increasing trial complexity within IRC -- for a number of reasons. It requires investigator-site compliance with collection and dissemination of scans to a reading facility, as well as afterwards storing all those scans. There's logistical considerations that still remain today with regard to missing scans, and the literature has been quoted that still today, 10 to 13 percent of those scans are missing, quality of scans, and variability of imaging techniques. Global trials can add additional challenges as digitized scans may not be available and techniques

may vary.

The IRC process itself requires the reading of scans from three trained radiology experts and requires specific training with a protocol development of a charter and application of RECIST criteria. Complexity of the investigator does determine progression since IRC will not receive follow-up scans, even if IRC has determined the patient responded, which then leads to complexity of trial analysis due to informative censoring that's needed. And there's additional site burden in already resource-stretched sites and complex clinical trial settings.

The next point is cost. When we collectively received feedback to see if we could get cost figures, some of the sponsors did provide that to us. IRC was seen, independent of the size of the sponsor, as costly. The average cost was estimated at $4500 to $7500 per patient. The total IRC review approximated 1 to $3 million, depending on trial size but per study, cost driven by the collection, storage and reading of scans. On top

of that, and not factored into these costs, are the operational and resource burdens to the sponsors, to the monitoring that needs to be done to manage and implement the IRC process. And one questioned that if it was not needed, could those cost savings be applied to other types of trials for potential new therapies.

We were asked if sponsors could comment on whether the sample audit approach could decrease burden and costs, as this has not been done, and obviously some of the sponsors did not have complete knowledge. Some do. As of today, some were part of the PhRMA working group working on this. We got some preliminary concerns. One was that the cost savings for conducting a sample audit may only achieve 20 to 30 percent cost savings, however, that's highly dependent on the size of the audit and whether all trials would need this to be performed.

There were some questions still on logistic burden of collecting all scans. So would sponsors still do this, or would they just need to collect

the scans to do the independent audit? There could still be sample size due to small sample size discordance, and then if the discordance is high enough, would it lead to a full IRC? Then the question is, if you did not collect all the scans, would you be doing a retrospective full IRC at that point in time? So concerns of the sponsors -- and obviously details may work this out. But they were concerned with the delay of access to patients if they had to retrospectively perform a full IRC late in the filing or NDA process.

With regard to regulatory considerations of the investigator sample audit, proposals for clarity or next steps, depending on the outcome today, are needed. The clarity from the agency of the investigator assessment will in fact be primary endpoint versus IRC. Where is the audit going to be placed within? Is that a secondary endpoint or another type of checking of the outcome?

We'd like FDA to consider, and the ODAC members to discuss, whether going forward -- once it's agreed to, and if it is, whether an updated

guidance on endpoints is needed, a white paper or guidance on the criteria, to use for a sample audit, the timing of that and how to conduct. We heard two different ways to conduct this audit. The question is whether it will be allowed flexibility or will there be one recommended way. We encourage the agency to publish their analyses in a peer-reviewed journal that was presented this morning. And most important, encourage the agency and all sponsors to have dialogue during development, especially through pivotal-trial discussions and SPA process, to have mutual understanding about predefined criteria for if and when an audit is needed. The last point is encouragement of FDA to speak to other health authorities because as many companies run global clinical trials, if there are changes recommended today, we would want to see if we could align with EMEA and Health Canada at least.

The second point is criteria, as I know it will be discussed today. But the sponsors would want, going forward, predefined criteria for when

an investigator sample audit is requested. Collective opinion was it should not be needed for double-blind trials if safety does not break the blind; should be strongly considered for open-label, randomized trials. We'd like the determination of sample size. And again, the point of whether all scans are needed to be collected or just for the audit, depending on if FDA sees that a full audit may need to be conducted, depending on outcome of the audit. So the concerns are basically outlined here, again, with not a lot of knowledge from the sponsors about the sample audit procedure.

So in summary, both IRC and investigator assessments are different ways to evaluate PFS. They have different strengths and weaknesses. But if the overall study outcome is comparable, despite patient level discordance, we think that should be the focus. Investigator assessment should be considered the primary endpoint in randomized PFS trials, as bias appears to be controlled through the published literature and the additional

clinical information relevant to the totality of the patient assessment is very important.

IRC does increase burden and does increase cost and complexity, possibly without adding substantial value at this point in time. The sample audit approach should be used judiciously with clear predefined rules for use. Thank you.

DR. SEKERES: Very good. Thank you very much.

We are running a few minutes early, and we are scheduled to break for lunch in about 45 minutes. I'm going to ask the committee, do you need a break, or can you hold out for 45 minutes?

I'm not seeing too many --

DR. WILSON: Why don't we take a short break?

(Laughter)

DR. SEKERES: We'll take the former chair's prerogative --

(Laughter)

DR. SEKERES: -- and go for a 10-minute break. Please come back promptly in 10 minutes.

(Whereupon, a recess was taken.)

Clarifying Questions from Committee

DR. SEKERES: Can I ask everybody to please take your seats, again?

So I thought what I would do to get discussions going is I'm going to read the questions at hand for us to discuss, just to remind us again how we should stay focused. And I'm going to try to summarize a little bit what we've heard already this morning. As everybody on the committee is jockeying to raise his or her hand, so that Caleb will recognize you, please nod or wave to Caleb, and he'll write your name down on a list, and we will go in order. And as a reminder, the process here is to speak only when recognized by me.

So the first topic for discussion is, given the information provided on random sample-based audit strategies, the variability in radiographic measurement, and logistical considerations, please discuss whether the current practice of complete-case IRC review of all patients should be

replaced by a random sample-based IRC audit. The second discussion point will be to discuss situations where a random sample-based IRC audit may not be appropriate.

Now, I promised FDA I would emphasize the point that we are not here to discuss progression-free survival as an endpoint. We need to stay focused on the topic at hand. And what I've heard today is the following.

Why discuss this at all? Well, it would reduce the cost and burden on the clinical trial investigators, avoid some of the missing data issues, and essentially streamline the process. Independent radiologic review, or IRC, of scans may lead to a greater than 30 percent disagreement at the patient level between the investigator and independent reviewer assessments and/or among independent reviewers themselves, but there is agreement between investigator and independent radiologic review on relative PFS treatment effects despite this.

There is an inherent measurement error that

exists in the reading of radiographic scans and disagreements between readers at the patient level, which are commonly observed. However, regulatory considerations are based on the relative treatment effect at the population level. In particular, when the FDA conducted a meta-analysis, there was a high degree of correlation between investigator and IRC-determined PFS treatment effects as measured by hazard ratios, with an R of .954. We heard today two different proposals for auditing strategies from Drs. Dodd and Amit. We heard about variability in CT tumor measurements from a representative from Duke. And we heard about an industry perspective.

So we'll get started first with Dr. Liebmann.

DR. LIEBMANN: I have a couple of questions that I wanted to address to Dr. Amit on his presentation. On his slide number 11, which showed the correlation between differential discordance and differences in hazard ratio, I have two questions. The first is, on the overlay in the red

dots, of the actual clinical trials, obviously one of those seems to be outside of the no-bias zone.

Did you actually look at any of these individual trials to see if they in fact map to your results?

DR. AMIT: Can you clarify what you mean by mapped to our results?

DR. LIEBMANN: So specifically, it appears that there's a point -- X axis, 0.2; Y axis, 1.0 -- where you have an actual trial overlaying it that looks to be outside most of the no bias. And so did you actually look at that trial and see was it flawed in some way?

DR. AMIT: That I believe was one of the smaller phase 2 trials that we looked at, so I think there was a lot of variability around it to begin with. And so, yes, I think we didn't look in a lot of detail, but, I mean, you would expect -- you wouldn't expect perfect agreements. You would expect to see some trials where the differential discordance might be pretty big, but the hazard ratios might be similar or vice versa,

just from the type of variability, that you don't have a perfect relationship.

DR. LIEBMANN: And also with these two plots, to a non-statistician, it looks like there's a fair amount of overlap between the plot on the left and the plot on the right, between the bias and the no bias. How does that factor into the model that you propose, then, for generating audit size and triggering an audit?

DR. AMIT: Right. So I think there's some overlay. Obviously, what would trigger an audit are points in the right quadrant there, the top-right quadrant. What would trigger a full case review from an audit I guess would be points from the top-right quadrant. And you can see quite a few of those in the bias case, and you see much less of those in the non-bias case.

DR. LIEBMANN: Although it certainly seems like a fair number of the bias case would be well within the no-bias range as well. Is that accurate?

DR. AMIT: Right. And that is a sense of

what we're calling sensitivity. And when you actually look at the sensitivity based on that simulated data, it's about 10 percent of the cases where you would miss bias that would actually be present.

DR. LIEBMANN: And so that gets to what I think would be my final question, which is, one of the big discussions seemed to be -- in looking at your audit methodology and the proposed NCI audit methodology -- the limit on the number of cases that your methodology appears to include. And so how much does sensitivity affect the number of cases?

So if you change your sensitivity to, say, 95 percent rather than 90 percent, now what would your upper limit of cases be? It's presumably not going to be stuck at 160 or 200 or whatever.

DR. AMIT: I guess it very much depends on what trade-off you're willing to make with specificity. So you could, by defining the threshold value at the right level, still have an audit of 100 to 150, but you would have a much

lower threshold value. And then you would be proceeding to a full case review much more often when no bias was present.

So I would say if you wanted 95 percent sensitivity, you probably would want to increase the sample size a bit in order to get better specificity.

DR. SEKERES: Dr. Armstrong?

DR. ARMSTRONG: For Dr. Amit, your slide, slide A-6, the correlation curve, while it's pretty close, it looks like the error seems to be -- or the difference seems to be that the independent review is calling a higher hazard ratio than the investigator review. I realize it's not very different, but if you look at the red line, most of it means looking at -- it's a higher hazard ratio for the independent review than the local review. Correct?

DR. AMIT: Yes. And that's not an atypical -- you would almost expect that due to the informative censoring. That's not suggestive of any bias. I mean, they're still very much

clustered along the lines. But when you have that informative censoring, when the scans are no longer available once the investigator's called progression, you would expect typically a slightly higher hazard ratio on the IRC. And so I think that's probably what's explaining that phenomenon.

DR. ARMSTRONG: That gets to my second issue, which I think Dr. Dinella had brought up, which there seems to be a bias for the late discrepancy rate because you're going to have more study scans when the investigator is not calling progression but the independent reviewer is, because the independent review's not done in real time, so the investigator is still treating the patient; they haven't called them.

So unlike the early ones, you aren't going to have study-related scans later on, and so how does that affect this bias issue? And I don't know if you or Dr. Dinella want to address that.

DR. AMIT: I can start I guess.

DR. ARMSTRONG: Okay.

DR. AMIT: So I would say I wouldn't read

too much into the terminology between late discrepancy rate and early discrepancy rate. I think most of the discrepancies that you tend to see in a trial are concentrated on the investigator calling progression and then the IRC not concurring with that evaluation of progression. And the reason that happens is you just don't have the subsequent scans. And you can -- I mean, the late discrepancy rate picks it up, and then the converse, where you have the IRC calling earlier than the investigator. But I think in terms of detecting bias, the late discrepancy rate probably is a bit of a more sensitive measure, but not for the reasons of how the study is kind of executed.

DR. ARMSTRONG: One of the issues that nobody addressed was the question about whether there's some benefit to actually having a higher rate of review early on in a large trial so that you can actually be looking at whether there's some characteristic of the study, presumably of the treatment or the patient population, that's leading to a higher discrepancy rate; and that that could

then inform how much you need -- you know, the audit rate I guess I would say. And that if you had high concordance early on, you could do -- so basically like a self-adjusting -- I'm trying to think of what the word is.

Nobody's talked about that. Is there any reason why that would be -- it seems like that would be a useful -- first of all, you could establish that you aren't seeing higher than expected discrepancy rates. You could even maybe do it based on sites or countries, places where you might have more concern about radiographic review.

DR. AMIT: So I'll speak to that quickly and then see if others want to jump in. I mean, we have considered trying to do that. Obviously, in our procedure, we need to have knowledge of treatment assignment and compare between the arms. So you'd have to do that through an IDMC, and the sponsor would have to remain blinded. We have considered it, but we really haven't developed that thought process too far. Of course, the other consideration there, it's not a random sample

anymore. You're sampling the first set of patients with some assumption that --

DR. ARMSTRONG: That the later patients are going to be equivalent, yes.

DR. AMIT: -- the last set of patients will be similar.

DR. SEKERES: I think FDA has a comment.

DR. SRIDHARA: The currently proposed methodologies are for doing the random sample after the study is done. The method that you're seeing, as Dr. Amit pointed out, will not be a random sample from the whole population, then, and you're just looking at that. It could be perhaps used for monitoring how the study is going, and you may want to correct a trend, a particular site, or whatever is necessary.

But overall, the results that we are seeing, we don't think that there is no discrepancy at all. There is discrepancy in all of the trials that we have reviewed. But when you are looking at the treatment effect, then somehow this over-read or an under-read, or whatever, they seem to balance out,

it appears. And when you are looking at the treatment effect between the control and the treatment arm, it seems to not bother us as much.

DR. ARMSTRONG: But I guess the issue -- you're right. I mean, the data we've been presented is that the hazard ratios ultimately end up being pretty close. But if there are cases where the hazard ratios aren't close, where there's some effect that's causing a difference in the read, it seems like you would want to know that.

DR. SRIDHARA: So that was the example that was presented. The carcinoid example was the --

DR. ARMSTRONG: Right.

DR. SRIDHARA: -- only one where we have clearly seen that we arrived at different decisions. And that's what would bother us if it is that different. We do sometimes that the hazard ratio by investigator may be .5, whereas by independent review, it could be .6 and vice versa, but the end decision is there.

DR. ARMSTRONG: Right.

DR. SRIDHARA: So if we are talking about

estimating the effect size itself, then it's a different issue.

DR. SEKERES: So I just want to play off of something Dr. Armstrong just was discussing. In my mind, the aspect of a learning curve to reading these scans is what I would call some real-time dynamics of trial conduction. And the question I think you were asking is was there an effort to look at earlier reads as opposed to later reads and see if that affected the clinical irrelevance factor or the differential discordance from the two methods.

A similar type of real-time dynamics question I would ask is was there an effort within the meta-analyses that were conducted to pull out trials where you may have had faster-growing tumors as opposed to slower-growing tumors; so where PFS was looked at in patients who were multiply refractory as opposed to patients who were initially presenting with metastatic cancer. And I throw that out there to be answered either by FDA, Dr. Dodd, or Dr. Amit.

DR. ZHANG: So let me start first. Within the FDA analyses, we also looked at subgroups that were not presented. We looked at different lines of therapy, so first versus subsequent and also maintenance. There were two maintenance studies. And the reason why we didn't present them is because the results were not any different. There were no differences between the subgroups with respect to lines of therapy.

DR. SEKERES: And again, I'm going to just repeat Dr. Armstrong's question. What about looking at assessments earlier in a trial as opposed to later in a trial? Did that appear to affect either the clinical irrelevance factor or differential discordance rates?

DR. ZHANG: So to that point, we also looked at subgroup analyses based on trials that were submitted based on interim results versus final analysis results. And the same thing, the results were not presented because there were no differences that showed up between those two subgroups.

DR. SEKERES: I saw Dr. Dodd rise briefly. Do you concur with that?

DR. DODD: Yes. I just wanted to add, I mean, I think there's -- it sounds like we're mixing potentially an education component, which is feedback to the local site radiologist with the endpoint evaluation. I think we need to keep those two separate. And I do think there are methods for going about, in a more adaptive way, performing the audit strategy, but to date we haven't fully evaluated that or even designed something like that. But I think that would be an interesting next direction.

One thing we have to be careful of when we do that is that we don't disturb the blind that the central reviewer radiologists have, which would mean that we need to wait until some of the patients in the trial are administratively censored so that we can mix in some that do not have progression.

DR. SEKERES: We have a question from Dr. Harrington by phone.

DR. HARRINGTON: I have a couple of questions, and I'll ask them here in a batch, and then take the answers listening because I know how hard it is to conduct a dialogue by phone in these settings.

First, I wanted to thank the speakers on the work that's been done to put some analytic effort into a problem that's been a vexing one for a long time. The work is very, very nice. It probably has a ways to go to mature, but it's a great start. So first a question probably just for the FDA, although it's true for all the presentations, I want to confirm something.

Raji, on your slide, slide 6, your meta-analysis, the trials that you looked at, all the phase 3 registration in solid tumors with PFS from 2005 to the present, was it a true meta-analysis?

DR. SRIDHARA: We did do the meta-analysis also, but the figures that we showed were based on individual trials.

DR. HARRINGTON: Sure. So I guess --

DR. SRIDHARA: It is all phase 3 trials. We

did not include any phase 2 trials.

DR. HARRINGTON: Okay, great. I mean, it's an important point because when we look at data like this, either in your presentation, or the one from the NCI group, or the one from PhRMA, we need to know that the trials being represented are representative of the population of trials that we will be seeing ultimately for regulatory approval.

Second question is this very, very difficult idea of evaluation bias and this presumption that the investigators may be subject to evaluation bias because they know the patient's on a clinical trial, and perhaps they know in an unblinded trial whether they're on the experimental or the control arm. It's also possible of course, as someone mentioned toward the end, that investigators are calling progression because they have a full set of clinical information in front of them in addition to the scans.

So my question is whether -- since we all care about what happens in the clinic here -- the FDA or anybody else has looked to see whether

evaluation in trial settings by investigator is really substantially different than off trial? In other words, is it really an evaluation bias or is it much more likely that what you might be seeing is what happens when a clinician integrates the information across a clinical picture in addition to those scans? Any postmarketing studies help with that?

DR. SRIDHARA: I don't believe anybody has done a postmarketing study of that aspect. We have had a couple of applications where clinical progression has been included in the assessment of progression itself. However, here we have looked purely at the radiological progression since IRC looks at only radiological progression. So we did not include the clinical aspect for this purpose.

DR. HARRINGTON: Okay. Thanks. At one level it's a technical point, and on the other level it is an important point because, in fact, what we all care about is how do these treatments perform in the clinic once they're approved; what is the population progression times of trial as

opposed to how they might be different on trial.

Then one last question, and I think this one's probably for Dr. Sullivan. One of the things that's in the background here and it's been mentioned a couple of times is that measurement error, just pure measurement error, can bias results toward the null so that treatments might look to be not quite so good as they would if you got perfect measurements in some parallel universe of what's going on with the tumor.

My question is whether the technology that's used by the IRC, by the independent committee, is essentially equivalent to what's being used in clinics either on trial or off trial. Is it pretty uniformly applied so that the measurement error you might see by independent review is roughly comparable to the measurement error that's going to happen by the investigators with their scanning equipment or off trial?

DR. SULLIVAN: I'm not actually aware of any data about that, and I did try to look at that before preparing my presentation. So I don't know

any data. But I think for IRC, for central review, they would use a standardized method, a standardized software, so all the measurements would be made using that software by the readers. So I think the variability would be somewhat less, but I don't know of any data to really substantiate that.

DR. HARRINGTON: That's relevant as well because, in fact, in the clinic, progression would be determined by clinical investigators and equipment as opposed to what might be used by IRC.

All right. Thanks. I'll go back to listening. Thank you.

DR. SEKERES: Thank you, Dr. Harrington.

Dr. D'Agostino?

DR. D'AGOSTINO: I have a few comments and questions. I've lived through a lot of adjudication and what have you, and I'm impressed, and I want to congratulate all the speakers for their presentations. I'm impressed by the correlations, but it's sort of after the fact. I mean, all of the data sets were under the context

where there was going to be the IRC looking at the data, and it sets people up in terms of knowing that their data is going to be evaluated by some other group and makes them a lot sharper and careful in the presentation they give. So I'm concerned about we don't know about that.

Let me just rattle my little questions or comments. The other is that there could be big implications. And we sort of hinted at it in the conduct of the trial with, say, the interim analysis and increasing sample sizes and adaptive strategies and what have you, and has the FDA and industry sort of thought that out. When you know there's going to be an adjudication -- I've not lived in trials where the investigator's call is the one that one runs for. And when we have -- I've seen lots of cases, as we do different trials, that the adjudications don't look the same as the investigators. And a lot of the interim analyses is going to be -- if you're going to hold off the adjudication until the very end, a lot of the interim analyses now may change in terms of

what's given, and it may make a difference.

Another comment I have is, if I hear correctly, you've been doing random samples -- Professor Dodd's material, Dr. Dodd's, and the FDA, you've been doing sort of random sampling of the subjects. But in a normal study, you'd have large centers, small centers, and you'd have to sort of make sure that your adjudication, that your IRC and so forth, picked up or was looking at cases from small centers in addition. So is your strategy, in terms of how you're going to do this audit, building into account that it's not just a random sample of all the cases, but it has to be a representation of all the sites? And that may change quite a bit the strategy of doing it.

Lastly, I just don't see where one is drawing the line in terms of here we need it, here we don't need it. And I'm wondering what kind of a monster will be released by saying, yes, we can settle with an audit. Maybe the FDA can address those questions and anybody else who wants to jump

in.

DR. SRIDHARA: So those are excellent concerns or questions that you have. In general, we have discouraged doing interim analysis on PFS. Any interim analysis with any endpoint, we all know that the estimates are old estimates of the treatment effect at that point. So the advice that we have been giving to the sponsors is that they should come to us with final PFS analysis.

Having said that, we do occasionally see very huge treatment effect sizes where they will come with interim analysis itself. I'm not sure that adaptive design and some of these are going to be used in this kind of setting that we are talking about. But let's say they use adaptive designs, or whatever be it, that even increase the sample size. If we think about the NCI methodology, it is after the study is done and all the patients are accrued, and you have an assessment by the investigator that there is a treatment effect. And only then do you think about going and doing the audit. So it is after the study has enrolled all the patients. And

the study aspect of it, even if there was an adaptive design, has been all taken care of. That's in my mind.

Regarding the random sample itself, we could consider a stratified random sample maybe. None of this has been worked out. The question whether it's first of all a good idea to do this and your other question about the examples that we presented, they were 100 percent IRC. So is it because of that that the investigator's assessments were better? And if we didn't have that, would it be different? And that's where we feel that we have to have a random sample-based audit, and we cannot go with totally investigator-determined PFS just yet. Maybe future studies may let us know more about it, and we may be more comfortable using just an investigator PFS. But at this time, it is like having a traffic police somewhere standing, and so how this audit may happen, hopefully that will control some of the things.

DR. PAZDUR: I like the IRS audit better, example.

(Laughter)

DR. SEKERES: I think you're about the only one who likes the IRS audit.

Did FDA have anything else to say in response?

DR. SRIDHARA: That's about all that we have. But as I mentioned in my presentation, today we have these two methodologies, but in the future we could have other audit methodologies that can be proposed. We are not set on it has to be one of these. But basically we can have this, and we have to figure out how we can do the sampling. There were some suggestions of before the data cut-off date -- let's say a month before that, or something like that -- have a random sample identified and tell the sites to be ready with the scans for those patients in case needed for IRC audit.

The point is if there is very minimal improvement that you see by investigator PFS, then there's no need to go and do any of the audits because the study is really not showing any clinically meaningful benefit. So you avoid doing

totally in such cases.

DR. SEKERES: Okay. Thank you.

Ms. Mayer.

MS. MAYER: Yes. This question and comment is for Dr. Dodd, I think, referring to your slide 9, which concerns clinical irrelevance factor. As I understood it from the presentation, you're assuming or assigning a clinical irrelevance factor, CIF, prior to the analysis itself, but in my understanding, what is considered an acceptable median progression-free survival benefit is something that's determined at later points in the process and in fact may even be a topic of discussion at this advisory committee. So I'm wondering, in practice, how that would work.

Then secondly, does CIF take measurement variability and measurement error into account? The example you gave, which I realize was just for the purposes of presenting the idea, was looking at a one-month difference. And I think we've heard enough about measurement variability to understand that that's not a meaningful difference in this

context. So I'd just like to hear a little more about that.

DR. DODD: Okay. Thank you for the excellent question. The first question was how do you select the clinical irrelevance factor, and that should be done prior to conducting the central review. And I think you can do it prior to conducting the central review and at the beginning of the study when you're planning the contingencies for performing a central review. And for any given cancer, we have some idea what the median progression-free survival time should be under the controlled treatment. And therefore -- I mean, it's similar to setting a non-inferiority bound, which is always a difficult thing to do, and we all scratch our heads about, well, is this big enough, is this too small? And I would imagine those discussions would follow along similarly to setting up a non-inferiority bound, but it can be done. Again, it's a clinical decision, not a statistical decision. But the two sites have to work together.

The other question about the measurement

variability I think is a very interesting one. The measurement variability here is different from what we typically -- and I use the term "measurement variability" because radiologists don't like to think they make errors, so I've been well-trained. The measurement variability is in the time-to-event endpoint. And we're not -- in typical statistical literature on measurement error, the attenuation in the effect comes from the measurement variability and in the covariate X. And this is a different setting which has not been as well studied as the standard measurement error setting.

So it's not clear what the measurement variability -- how much it changes the effect size here, but it is something to consider. I think the informative censoring is also something to consider because there are things that will tend to attenuate the effect. And, therefore, if you set too high of a bar for the clinical irrelevance factor, you may not attain the significance that you're looking for, but it may be because this measurement variability and informative censoring

bias is floating around.

DR. SEKERES: Okay. Next, Dr. Fojo.

DR. FOJO: So I just had two questions. One, Dr. Dinella, maybe you could just clarify. You had the preliminary concerns, and you said that conducting a sample audit may only achieve a 20 to 30 percent cost savings. How did you come to that? Was that an opinion, and why was that opinion voiced?

DR. DINELLA: Yes. I think that slide, as I tried to caveat it -- but the sponsors -- this was collective feedback. The sponsors, to this point, don't have real information about the cost savings. So the only thing that was projected to the sponsors is the point that it's not a one to one. So doing a sample audit, it may not be substantial savings, but, at the same token, there will be some savings. The numbers, we don't have numbers because this hasn't been done before.

Your question is, to what trials. So depending on the criteria, if you don't have to do it for double-blind trials, well, that's savings in

itself. If you had to do it for some trials, what's the magnitude of the sample size? And I think we've heard two different proposals. So, again, it's projected, it's anecdotal, but not based on fact.

DR. FOJO: Okay. And, Dr. Zhang, you picked the carcinoid trial. That was the trial that at the last minute was pulled from the consideration by the ODAC, wasn't it? I mean, so that trial had a lot of problems, not just the problems that you address here. It just kind of stands out.

Are there any other trials that you found that sort of were indicative of problematic trials? I mean, even in this trial, independent review didn't correct the problems; it just, in fact, created more problems. Right? So where did independent review really help? The data seems to say nowhere. And I thought you were using this more to show -- actually, what this ends up showing to me is that audits aren't going to solve that either. So how do you see that?

DR. ZHANG: Right. The main point of using

that trial as a case study is to be able to differentiate a little bit between the two audit methods and to see whether or not -- or how the two audit methods would perform for a trial in which there was very high confidence that there was something wrong and that there was evaluation bias present.

You bring up a good point that in such cases in which you go to the full audit and you get these discrepancies, still what do you do from then on? And I think that's not where the savings comes in, certainly, with the audit. And that's why I had presented the sarcoma study as well to kind of serve as a counterpoint to the carcinoid trial.

So with respect to the carcinoid trial, really, in that case, the actual trial itself was pulled. And so I think that just goes to show that that audit really didn't confirm the effect by the investigator. And so the data perhaps is just not good enough to be able to tell us anything about the treatment. So, again, that's not where the savings is. So the illustration with that trial is

really to show whether or not the two audit methods could pick it up that there was evaluation bias. And in this particular case, the NCI method was able to do so, whereas the PhRMA method fell a little bit short.

DR. FOJO: Okay. And maybe just a general question that I was just interested in. I mean, I was surprised, Dr. Sridhara, how good the data was in terms of response rate, which people usually say, oh, it's higher usually with the investigator than with the independent audit, but you had showed pretty good correlations between those as well.

Did you notice -- do the independent reviewers measure a different quantity of tumor? Are they measuring bigger tumors that then have to shrink more before you get a response, or did anybody, in looking at data that they had, find a difference between how much RECIST -- what the RECIST quantity was of the independent versus the local assessment?

DR. SRIDHARA: We didn't go into that granularity, looking at the tumor sizes

individually and how it was affecting the investigator versus the independent review, because we have seen that there is discrepancy anyway. So whether it is because they were choosing different lesions or what not, anyway, there are discrepancies. So that we see across all trials.

With respect to response rate, what we saw was investigators were always calling more response than independent review. But, however, when you are looking at the relative treatment effect between the control and the treatment arm, it seemed like at least the data -- that's what we have -- is showing that the treatment effect was smaller by investigator compared to the independent review. So although the independent review was calling less responses in both arms, the overall treatment effect, the relative treatment effect, was larger there.

So that was a bit of a surprise for us, because, particularly in single-arm trials, we have seen that every time, as we saw here, too, the investigator response is always a much higher

response rate compared to the independent review response rate.

DR. SEKERES: Just a point of clarification from something Dr. Fojo just said. The R for hazard ratio, for PFS, was .954, but for response rate was approximately .7. We're considering this for hazard ratios for PFS, not response rate.

Is that correct?

DR. SRIDHARA: For the response rate, it was the odds ratio that we were using. And, yes, the correlation is not as high as PFS.

DR. SEKERES: So just to clarify, we're talking about this issue with respect to hazard ratio for PFS, not response rate.

DR. SRIDHARA: Say that again.

DR. FOJO: I mean, I think what you're asking is this is all about whether we need an independent to call PFS, not studies for response rate as the endpoint. Correct?

DR. SEKERES: I think we want the discussion to focus on PFS, not response rate.

DR. SRIDHARA: Yes.

DR. FOJO: Yes.

DR. SEKERES: Because you did make the comment that it was good for both, but it wasn't quite as good for response rate, actually.

DR. FOJO: Right.

DR. SEKERES: And I think we're focusing on hazard ratios here.

We're going to take one more question from the panel, then we're going to break for lunch. But please be reassured we have a list here that Caleb is going to lock in a safe during lunch, and we're going to get back to it and go in order.

So, Dr. Wilson.

DR. WILSON: So this question is for FDA. And if in fact you do adopt this random sample procedure irrespective of the methodology, your meta-analysis is -- I mean, I think Dr. D'Agostino already brought up the fact that you looked at trials that had already been approved, so I think you're more likely to have seen a concordance between the two. But there are certain tumor types where we know reading is very difficult, end stage

ovarian cancer, carcinoid. You really don't have adequate data for certain settings.

Is it your -- or maybe I haven't worked this out. Do you plan on -- if you do implement this -- applying this random sample to tumor settings in which you, a priori, perhaps know there will be more difficulty, or are you going to require that they have full auditing, and, therefore, as you get more information, perhaps decide that you will do this random sample method for them? That's my first question.

DR. SRIDHARA: So it is an excellent question, but we wanted to stay away from discussing on specific tumors here. And I think in some of those tumors where it is difficult to measure, we may not even have PFS as the endpoint or may not accept. So, yes, we will have those considerations while considering. And, Dr. Pazdur, may want to --

DR. PAZDUR: I think, as pointed out by several of the speakers, you have to take it on what tumor you're measuring and the difficulties.

And this may not be an approach that fits all patients. So if there is a great deal of difficulty, and the numbers of patients that are enrolled on a clinical trial also, that may come to bear in to the consideration of what should be the size of the audit or should it be 100 percent review of the X-rays.

DR. WILSON: That was exactly what my question was. I wanted to clarify that this isn't -- I understand this is a work in progress.

My second question is, there are tumors in which there's a combination of both measurements as well as biomarkers. I think prostate cancer would be one; germ cell cancers would be others. You didn't discuss anything that would integrate biomarkers within the solid tumors. Do you want to comment on that?

DR. PAZDUR: We really don't have data on that, so that's something that would have to be investigated. For the most part, we have not just -- in our prostate cancer trials have been looking for radiographic progression, not just

measurements of PSAs. And the germ cell tumors, we haven't an application on that. And it's such a very specific disease with very highly effective therapies. I think we'd really have to take a look at drugs that are coming into play there.

DR. WILSON: Right. I was thinking that usually in those settings, you wouldn't use one or the other. It's often integral, so it's more complicated.

My final question is you've stayed completely away from all of my tumors, which are hematologic. And I think that the RECIST criteria has not been applied to them yet. They are more complicated to measure because they're bidimensional measurements, and there are many complex biomarkers integrated there, too.

Do you want to just comment on the fact that you've stayed away, or is this just a topic for another setting?

DR. PAZDUR: Well, as pointed out by Raji in her presentation, one of the problems there, other factors come into play here other than just

radiographic measurements, including blood counts, physical examinations. So that's kind of a mixed bag. I think we would really have to take a look at those and do the same type of analyses on the specific tumors before we leap into that area, because these other factors come more into play.

DR. SRIDHARA: Also, there is a lot more heterogeneity in that disease, and so we don't know whether this will work there at all.

DR. WILSON: So just to clarify, then, we really are talking about this --

DR. PAZDUR: Solid tumors.

DR. WILSON: -- really pretty much to standard, solid tumors, et cetera.

DR. PAZDUR: Correct. At this time, radiological review of solid tumors. That's why we really wanted to emphasize this. And we only selected solid-tumor trials in our analyses.

DR. WILSON: Great. Thank you. That's it.

DR. SEKERES: Okay. Thank you, everybody. We will reconvene in this room in 45 minutes, at precisely 12:45, to get started again. Panel

members, please remember, there should be no discussion of the issue at hand during lunch amongst yourselves or with any member of the audience. Thanks.

(Whereupon, at 11:57 a.m., a luncheon recess was taken.)

A F T E R N O O N S E S S I O N

(12:47 p.m.)

Clarifying Questions to Committee (continued)

DR. SEKERES: Good afternoon, everybody. I think we're ready to get started again. On the schedule is the time for the open public hearing next, but I did want to continue with some of the questions we didn't reach from the morning session because I think sometimes when we're talking about the details of statistical analyses, the half-time deterioration is short. And I figured I better try to time this to as close as possible to the presentations we heard earlier.

The next person on the list who had a question was Dr. Logan.

DR. LOGAN: I had a couple of questions and clarifications mainly. The first one is we've been talking about the strong correlation between the hazard ratios for progression-free survival between the IRC and the investigator. Oftentimes at these meetings, we discuss the median time to progression-free survival. Does that strong

correlation hold also for the median progression-free survival or the difference in median progression-free survival?

Has the FDA looked at that?

DR. ZHANG: Yes. We did actually look at the median progression-free survival times as well. It's actually -- give me one second to find you the details.

Actually, Dr. Logan, would you mind coming back to me?

DR. LOGAN: Sure.

DR. ZHANG: I'll find it for you.

DR. LOGAN: I'll continue with my questions, and you can come back.

I had a question for Dr. Dodd. My question that I had, had to do with how the size of the audit was selected. Is this based -- is there a -- I know from looking at the paper, at least in the simulations, the method for selecting this audit size requires some knowledge about the correlation.

Is there some assumptions about that or is

that -- how is that done?

DR. DODD: Yes. So the audit size does require an estimate of the correlation between the central review hazard ratio and the local evaluation hazard ratio.

Is that the question you're asking?

DR. LOGAN: Yes.

DR. DODD: Yes. So in practice, in order to obtain an estimate of this, we typically would recommend taking an initial audit to estimate the hazard ratio or the correlation between those two hazard ratios. However, I think as we move forward, we may be able to avoid that step because we are getting a better idea of what types of correlations to expect. And so from Dr. Zhang's work, we now presumably have estimates on the correlations for 28 trials, and that may guide that. We might want to choose a conservative estimate based on the data that we've already collected. But one could go about it that way or go about it by getting a preliminary estimate from a small audit to estimate it.

DR. LOGAN: I mean, I guess just from a logistical standpoint, the need to do several steps, I guess, would be less appealing. So if we have that preliminary information, that would be helpful in applying the method.

Did you have the results?

DR. ZHANG: Yes, I do. So in our meta-analysis, the mean difference in the investigator and IRC median PFS, in months, in the control and experimental arm was .94 for the local evaluation and .8 for the IRC.

DR. LOGAN: That was the mean?

DR. ZHANG: Yes, the mean difference of the medians.

DR. LOGAN: Mean difference in the medians. Okay. Do you have the correlation between the differences?

DR. ZHANG: Between the medians, no. We didn't specifically do that.

DR. LOGAN: I had a question for Dr. Amit. In your slide A-17, you had some threshold values that you had described based on sensitivity and

specificity. The question I had is do those assume a particular trial size? Not the audit size. The sample size that's given there I assume is the audit size.

The question is what is the fraction of the -- what is the fraction of the audit size versus the trial size for those, and does the threshold depend on that?

DR. AMIT: The threshold does not depend on the fraction of the audit size relative to the trial size. The threshold depends particularly on the size of the audit. And it's chosen in a way to fix the sensitivity at 90 percent, and then based on increase in sample size is setting a higher sort of threshold -- a higher bar, if you will, so as to increase the specificity.

DR. LOGAN: So your definitions of sensitivity and specificity is based on the truth being whether the full -- where there's a discrepancy in the full population, right?

DR. AMIT: That's correct. It's based on --

DR. LOGAN: So those thresholds don't depend on the ratio of the fraction of the audit size?

DR. AMIT: They do not, no. They depend on, primarily, the size of the sample.

DR. LOGAN: Okay. And then my last question is for the FDA. I just wanted a clarification in looking at the results that were presented, slide 16. So you described applying the two methods to these trials. In both cases, you did simulations of the audits, of the random samples of the audits.

Is that correct?

DR. ZHANG: Yes, that's correct.

DR. LOGAN: Okay. So in terms of understanding the PhRMA method, do you have -- are those results based -- were the results consistent across all the samples of the audits or not? I guess I can't interpret the number for the full audit for the PhRMA method.

DR. ZHANG: Right. So with respect to the PhRMA method, we fixed the audit size to 160 patients, and that size was sampled 1,000 times.

And the results are an average of those 1,000 replicates. And then, depending on that, the threshold values, which is also an average, was taken and compared to, in this case, .075. So the differential discordance in the early discrepancy rate and the late discrepancy rate was compared to the threshold of .075. And if it was greater than that threshold, then, according to the method, a full audit would be recommended.

DR. LOGAN: So the 42 percent is averaged over both the number of trials where the investigator hazard ratio is below .5 and the 1,000 simulations?

DR. ZHANG: Exactly.

DR. LOGAN: I'm sorry. I just have one last question. I guess I'm a little concerned about slide 20, where the PhRMA method doesn't seem to be picking up the discrepancy that's been raised here. Do you have any idea of why that is? And also related to that, it says "full audit, no," but were you also simulating 1,000 times here? Do you have the percent of time that they went to no audit, to

no full audit?

DR. ZHANG: So they were all no audits because --

DR. LOGAN: They were all no audits.

DR. ZHANG: -- right. Essentially, the numbers presented here -- so the differential discordance that's shown for the EDR and the LDR are using 160 patients sampled 1,000 times, and then those 1,000 values for the differential discordance were averaged.

DR. LOGAN: Averaged.

DR. ZHANG: So that's what you're getting here with the .001 and the .01.

DR. LOGAN: But the variability was such that it never went above the threshold for triggering a full review?

DR. ZHANG: Right. Right.

DR. LOGAN: In those 1,000 times.

DR. ZHANG: Right.

DR. LOGAN: So do you have any sense of why that is? It's been raised that the carcinoid tumors are problematic, but --

DR. ZHANG: Sure. But I think one possibility -- and this obviously needs more evaluation. If you'll look at the discordance rates table, you'll see that the censoring status percentage discordance between the two treatment arms are actually quite discrepant. You have 38 percent versus 26 percent. And if you would recall from the 2x2 table of the PhRMA method, the formulas for the early and late discrepancy rates do not utilize the one cell that's designated D, which is where both the investigator and the IRC have censored the patient. So one possibility is that ignoring that information may have some impact in certain cases. It's one hypothesis for potential future evaluation.

DR. SEKERES: Okay. Dr. Steensma.

DR. STEENSMA: My questions have largely been asked and answered by other panelists, so I'll be pretty brief here. It is very difficult I think to consider this without considering the validity of PFS in certain sorts of circumstances, but I'll try to divorce things here.

I'm trying to get a sense of just how much of an outlier this carcinoid study was. We've all had concerns for many years that radiologic assessment may be problematic, and it's reassuring, the concordance that was seen, both in the data presented here and the data presented by Dr. Dodd. So my question to Dr. Sullivan is, are there other examples, besides the carcinoid study, of large trials in oncology, that may not have come to the regulatory agency, where a discrepancy or discordance did change assessment of treatment effect?

You cited several examples where even though there was a discordance, it didn't obscure treatment effect, which is reassuring. So I'm just wondering is this an anomaly.

DR. SULLIVAN: I don't know of any others. I haven't found any in the literature. And when that discussion occurred last year, I asked various people in the industry and in radiology if they knew of other examples, and don't know of any others.

DR. STEENSMA: Thanks. And then the other question was just for the agency biometricians. In the analysis, the meta-analysis of the 27 studies, is that carcinoid study in there on the plot or was that excluded?

DR. SRIDHARA: Yes, it is there.

DR. STEENSMA: Is it possible to point it out? I'm just wondering where it would have fit on your slide 7.

DR. ZHANG: So could I have the backup slide, number --

DR. STEENSMA: Number 9, I guess it would be.

DR. ZHANG: Number 4, please, Caleb.

So it's still a little bit hard to tell, but this is the same plot that was shown in the presentation, except that now there are labels for each of the points with the indications.

DR. SRIDHARA: But carcinoid would come under "other" so --

DR. ZHANG: Right.

DR. SRIDHARA: -- all the others are just

around the line as well. You can't really see it being totally off the line.

DR. PAZDUR: I think to answer your question about whether this represents kind of the universe of trials, I think it probably does, and let me tell you why. Because even if there was a difference between the investigator and the IRC, and one was positive, I almost guarantee you the sponsor would be coming to this body to argue why one of these readings was the correct reading, if they did have one of these readings at least positive, statistically positive.

We haven't -- since I see the trials for all of the three divisions here, I don't recall any that were not submitted because of a discrepancy between an investigator and an IRC reviewer. I'm pretty sure that there wouldn't be an argument, that, yes, this one represents the truth, so to speak.

DR. SEKERES: Okay. Dr. Balis.

DR. BALIS: Thank you. I think that if we are going to see investigator bias on a study, it's

most likely that we'll see it on the experimental arm of a randomized study, at least presumably. The FDA, when you did your analysis of the 27 studies, you presented that kind of as a whole with all the arms on those slides. Did you look at whether there was a difference in this degree or direction of discordance for experimental versus control arms on open-label versus blinded studies? Because theoretically, if there really is a bias, you would see it mostly on the open-label studies in the experimental arm, or at least the direction may be different.

DR. SRIDHARA: So the slide that I presented, slide number 8, it differentiates between the open-label and the blinded studies.

DR. BALIS: Right. But I'm talking about within those studies, you have an experimental arm and a control arm. So were there differences between the experimental and control arms on the studies, depending on whether they were blinded or open label, in terms of the degree or the direction of the discordance?

DR. SRIDHARA: No. There was no particular direction that we saw that stood out. No.

DR. PAZDUR: This may differ from what you're thinking, if there is cross-over, especially to the experimental arm that isn't available and people calling on the control arm earlier if they believe that the experimental arm is better; and especially if this therapy has been touted prematurely in the medical literature as having response rates, and a conventional therapy is very poor in this disease.

I've seen some tendencies of these -- I don't know if they were included, but some suggestions of this, where people were calling people on the control arm early progressors so they could go on to this magical therapy. Here again, that's one of the inherent problems with a cross-over design.

DR. BALIS: The other question I had -- you kind of alluded to this just a second ago when you were talking about what the sponsor brings. But if there's a trial that's done where there's a central

review, and you see the results of both, and there's a discordance, do you always assume that the central review is correct and the investigator interpretation is not?

DR. SRIDHARA: No. I don't think we can assume that. I think we will do a more thorough investigation and see if reasons can be explained why it is different and why we should believe one versus the other. The point is it's the same scans that both of them are reading. It's not that they're getting a different set of scans.

DR. BALIS: No. They're the same scans, but because of the way the criteria is written, they can look at different lesions on the same scans and measure them, or they could measure a different cut. There are lots of reasons that there could be a different analysis of the same scan.

DR. PAZDUR: There's no absolute truth in this. Obviously, there are biases that are present in both of the readings, especially with the censoring that can occur on the investigators. That's why we want it specified up front whether

it's going to be the investigator or the IRC PFS determination that will be used as the primary endpoint of the trial. So it should be prespecified before and not going back and saying, well, this is the correct one; this is not the correct one.

So what we are looking for in the statistical plans on all of these, especially when they go through, for example, a special protocol assessment, is a delineation of which of these endpoints are going to be the preferred endpoint or the primary endpoint of the trial, whether it's the investigator or the IRC.

DR. BALIS: So I wanted to make a comment because one of the things that's underlying, I think, the problem that we're trying to address today is in the way that we not necessarily measure these tumors, but what we do with the data, and that is that we categorize it. We take an absolute measurement, look at a change in tumor size, and then we convert it either to say it's a progressive disease or not progressive disease. So we take a

lot of continuous data, as Dr. Sullivan mentioned, and we convert it to a dichotomous endpoint. And I think we oftentimes create discordance by doing that.

The extreme example is a patient who comes in with a tumor that's 2.1 centimeters, which is considered measurable by RECIST. And the investigator at follow-up measures it as 2.6, which is a 24 percent increase. And the central reviewer measures it as 2.5, which is a 19 percent increase. And that patient is going to be off study and censored simply because of a 1 millimeter difference in the measurement because we have a bar that we set. If these data were analyzed continuously, those would be the same.

DR. PAZDUR: It's hard to analyze data continuously in the sense, but one of the, I think, central points here that we'd like to get across is that we would expect those same discrepancies to occur randomly in both arms of the study. And when you have random errors that occur between both arms of the study, it simply adds noise to the study,

and it's harder to demonstrate a superiority claim in a randomized study here.

That's why we're not so much concerned about this 30 -- and we admit to that throughout all of these presentations, that there's roughly about a 30 percent discrepancy rate, whether two radiologists read them, three radiologists read them, the investigator, a radiologist, reads it. And those go to the subjectivity of reading an X-ray. And that's the point that we're after. We're not so much interested in this 30 percent. We're interested, is there a bias that is here that really negates the trial. Because, really, if you have a bias that is present here, as we saw in the carcinoid trial, it renders the trial uninterpretable, and that's what we're really after here.

DR. SEKERES: Just to emphasize that point, I think the agency's been pretty clear about the fact that we're thinking about this on a population and not an individual person level, and that there will be some discordance on an individual patient.

But it's probably not going to be major, and it's going to be introduced random as classification bias, which isn't going to be systematic, which is what we're worried about.

It's also a point that I think Dr. Dodd mentioned about the performance characteristics of that strategy, depending, to some extent, on effect magnitude. If it's going to be a small effect, you're going to need a greater sample size to detect that effect as opposed to a larger effect, where you don't need as much.

Dr. Fingert?

DR. FINGERT: Thank you very much. Dr. D'Agostino earlier asked if industry had considered the consequences -- I'm paraphrasing please -- of how this might proceed forward. And I'd like to respond to that and also ask a question to some of the speakers.

So on this important topic, there has been a keen interest by people in industry, going back several years, in multiple groups in industry. And through bio organizations, I've been privileged to

participate in a recent roundtable discussion about oncology product development. And on this specific draft guidance, for example, there's been general consensus, very positive, for moving forward on this.

All were grateful to the agency and the participants to pursue what some people called the least-burdened principle or the least-burdened approach, as a general topic here. Some favored this as a preliminary step to future elimination or curtailment of all such central reviews. And some also saw value to retain an active role by the agency for consultation, depending on the registration trial that's being proposed.

I was really intrigued that some really did not see it as this simple formula to cut costs. Instead, they viewed it as, I think wisely, a need for basically reallocation of resources, the idea being that if we're going to reduce reliance on the IRC, then there's going to be greater reliance on the local evaluation, and what can we -- either as industry sponsors or collaboration with other

organizations, nonprofit organizations or NCI -- do to partner and to develop higher quality training and sharing experience, and then develop some answers to some of these multiple questions about operational aspects.

Now, some have volunteered to host or participate in open conferences that are coming up. BIO, for instance, was proposing a WebEx about this topic. And the PERI organization, the Pharmaceutical Education Research Institute, has an October 22nd conference about oncology trial designs. And they said they'd be happy to dedicate hours in that conference to this topic.

So it gets to my question now. My question is really to Dr. Dinella and some of the other speakers this morning. Would any of you see this as a follow-up? Do you have visions of this kind of follow-up being instrumental to help the guidance become a practical reality, and would you participate in things like this?

Dr. Dinella, could you respond?

DR. DINELLA: I think what you've raised is

the question over time, if we're going to investigator assessment, is how to increase the rigor of local evaluations, similar to the rigorous view of the IRC and how to do that. I think the onus -- and when you talk about reallocation of resource, the question to the sponsors is how to ensure that happens. And I think Dr. Sullivan has raised some valid points about some of the issues in different radiologic readings.

So whether that's through different types of -- whether it's publications, training, training courses, et cetera, if this goes broad-base outside three independent readers who are well-trained to IRCs, I think your point does balance something for us, to think about how to proactively do that.

I invite -- if there's any additional. Dr. Sullivan?

DR. DODD: Let me just add one thought that we had proposed a while back, which was if we can ensure blinding of the local evaluators, that would go one step further in that direction. At this point, I don't think we really know how many of

those evaluations are being done by a radiologist in their radiology suite, or the treating clinician is there assisting or pointing out this region; let's look at that. I don't think we really know how much potential bias there is in terms of stretching these lesions to call progression earlier, say, in the control arm.

DR. SEKERES: Dr. Shankar?

DR. SHANKAR: Thank you. I'd like to start with actually asking a clarification from Dr. Dodd. So in the methods that you've proposed that have been published, is the audit sample -- well, it's decided as you decide on what a clinically meaningful benefit is, correct?

DR. DODD: Yes.

DR. SHANKAR: And do you decide which cases would be in that audit? Are they pre-identified or does that happen at the end of the study?

DR. DODD: So it would be -- you would typically define it -- I mean, the way we have it set it up, at the end of the study, you would take a random sample. But one could modify that. One

could modify it to over-sample events in a case where you have a low event rate, or one could -- once enrollment is completed, one could very easily select that list of ones that you would sample.

DR. SHANKAR: So as you see it, one of the possibilities is that all the cases would still have to be collected and stored, in a sense, for potential future audits.

DR. DODD: Yes, right.

DR. SHANKAR: Thank you.

The other question I had is actually both for the FDA as well as Dr. Dinella. The first question is, what percentage of scans -- how much of the data are you actually seeing for these multiple time-point -- multiple radiographic endpoints, whether it's an independent review or a site evaluation? Do you see 90 percent of the data when these studies come in, or do you see 100 percent? What sort of data loss do you have in a clinical trial?

DR. SRIDHARA: So what we see -- the meta-

analysis that we have used for every one of the -- whatever the investigators read, they were all available for the independent reviewer as well, but they did not agree on some of them, whether it was progression or censor. We did the tumor measurements, not the scans. We don't get the scans.

DR. SHANKAR: Not the scans.

DR. SRIDHARA: No.

DR. SHANKAR: But when you do a central review, you would presumably need the scans, at the end of the study. If you had to identify an audit sample, you would actually need the scans, correct? To be able to do the review.

DR. SRIDHARA: See, those are the source data, and they would be at the sites, so whatever would be available for FDA inspection. But what is submitted in the application is simply the tumor measurements and not the scans.

DR. SHANKAR: Okay. Thanks. And, Dr. Dinella, you usually have 100 percent of the data, at least from your experience?

DR. DINELLA: Yes, on behalf of the sponsor's data, minimal is missing --

DR. SHANKAR: Minimal.

DR. DINELLA: -- at the point and time of submission.

DR. SHANKAR: And the other question I had, for the studies with the PFS endpoint, where it's a site review, do you have a sense of how much is read by a radiologist as opposed to a clinical research associate or an oncologist? Do you have any such data either at the FDA or from the company, as Dr. Dinella presents?

DR. SRIDHARA: We don't get that kind of granularity. We assume that it is the local or investigator site read, and it could be any of them, and if they have the IRC. So probably this will be the practice, and then some of them, it will be radiologist, and some of them, they may not be. But overall, the results are what we have.

DR. MURGO: And unless it's prespecified, one would anticipate that it would vary from institution to institution, site from site, as to

why does the read -- whether it's the investigator or whether it's the radiologist.

DR. SHANKAR: Right. And it might even vary on a site from day to day, but I was just wondering if you have a sense of what percentages are read by whom. And a particular patient, it wouldn't be implausible for site reads to have one set of scans read by one particular -- a radiologist, the next time a CRA, or whatever.

DR. SEKERES: Okay. Our final question following up from the morning from Dr. Wozniak.

DR. WOZNIAK: Thank you. I was trying to think a little practically. From reviewing the data and hearing all this, it doesn't seem to me that there is all that much bias. The other thing is that despite differences in measurements, the concordance with hazard ratios in terms of outcome seems to be similar. So, to me, that would indicate that probably a number of these trials actually don't need independent review. And I was just wondering if the FDA, or anyone else who wants to chime in, if you've actually sort of looked at

this and figured out which trials you think you really need to do this in.

I mean, for instance, you might not need to do it in a trial where you're studying a tumor that's very malignant, very aggressive, because there's not going to be a big issue about progression, but something like carcinoid, maybe it's important. Or maybe in addition to tumor type, there may be other issues as well. Maybe completely blinded trials, which are somewhat unusual in oncology, may not need to be -- and that was mentioned actually in the reading material.

So I was just wondering if you've actually thought about this in eliminating some of the trials that you really don't need to do it in. And in the ones you feel that it might be useful, to look at which ones might need all the patients reviewed by an independent panel versus just random sampling.

DR. PAZDUR: The answer is yes. Okay. But let me address this. It's a very complicated question. First of all, as was pointed out by many

of the presenters, what we're really interested in is the control of bias. And, generally, when one takes a look at other therapeutic areas, bias is controlled by blinding. One of the problems that we have in oncology is that few of our trials can be really effectively blinded. Even in trials that state that they are blinded, the differences in toxicity frequently unmask that blinding. So there is always that possibility of an introduction of bias. And from a regulatory point of view, it's very important for us to have some estimation or understanding is bias creeping into that trial.

This is particularly bothersome in oncology, where we are approving drugs uniformly on the basis of one clinical trial. Let me remind you, in other therapeutic areas for the approval of new molecular entities, two clinical trials are routinely used. So, really, when we're dealing with one clinical trial, we really would like to make sure that a bias has not been introduced.

Can we take a look at other endpoints such as response rate which might corroborate that?

Yes, and that might give us some comfort. But generally those other endpoints are observed after the clinical trial is near completion.

So those are some of the problems that we grapple with here, is we have one trial. We have to make sure that that is a real endpoint if that is the primary endpoint. Is that endpoint corroborated by other evidence? Yes, that may come into play here, but I think it's important. When one takes a look at doing away with some attempt to really measure this bias or assess this bias, it is somewhat bothersome.

Remember, it goes back to the former discussion that we had. Some of the reasons that we haven't seen bias creeping into trials, really, is because we've had this procedure in place. And that's an unanswerable question. And I was joking when I mentioned the IRS, but we had this discussion amongst ourselves. If you just did away with -- and just announced at one time, well, everybody, it's an honesty system here, we're not going to have any audits, what would be the

compliance of people paying their taxes, et cetera? I'm not trying to use that as a direct comparison here, but it does point out to some of the problems that could creep in when you don't have any assessment of the introduction of bias in a clinical trial.

So the major issue here, few of the trials can be blinded effectively. Okay. We are dealing with a subjective endpoint. We have one trial. We need to have some idea, at least initially, if we move away from 100 percent review of X-rays, and some type of attempt to address this issue and to measure it.

Open Public Hearing

DR. SEKERES: Okay. I'd like to thank everyone for their insightful questions. We're going to move on to the open public hearing. And I do want to think also the people who will be speaking in the open public hearing for their patience, as we're running a little bit late. But I assure everyone, we will finish by 3 o'clock today.

Both the Food and Drug Administration and the public believe in a transparent process for information-gathering and decision-making. To ensure such transparency at the open public hearing session of the advisory committee meeting, FDA believes that it is important to understand the context of an individual's presentation.

For this reason, FDA encourages you, the opening public hearing speaker, at the beginning of your written or oral statement, to advise the committee of any financial relationships that you may have with the sponsor, its product, and, if known, its direct competitors. For example, this financial information may include the sponsor's payment of your travel, lodging, or other expenses in connection with your attendance at this meeting. Likewise, FDA encourages you at the beginning of your presentation to advise the committee if you do not have any such financial relationships. If you choose not to address this issue of financial relationships at the beginning of your statement, it will not preclude you from speaking.

The FDA and this committee place great importance in the open public hearing process. The insights and comments provided can help the agency and this committee in their consideration of the issues before them. That said, in many instances and for many topics, there will be a variety of opinions. One of our goals today is for this open public hearing to be conducted in a fair and open way where every participant is listened to carefully and treated with dignity, courtesy and respect. Therefore, please speak only when recognized by the chair. Thank you for the cooperation.

We'd like to invite the first speaker.

DR. PATT: Thank you. I'd like to thank the committee and FDA for the opportunity to speak today. By way of introduction, I'm a radiologist. My background, for the last 20 years, I've been involved in various levels of oncology clinical trials as both a co- or sub-investigator. While I was in academics, I have been a site reader and a central reviewer. In the last six years, I've been

involved in either design and/or implementation of the imaging for over 150 oncology trials.

What I want to speak a little bit about today is really some of the operational issues related to relying more on site reads for support of primary endpoints. There have been some excellent statistical arguments supporting the audit methodology today, and I'm not here to either support that or refute that, but just to -- as someone who has been in the trenches and am in the trenches of involvement in site reads -- bring to light some of the operational issues that may impact the use of relying more and more on these site-read processes.

One of the speakers this morning talked about the goal of the potential audit process is to increase the confidence, the integrity of the trial and trial endpoints. So what I really want to talk about is site-read data integrity here. I'm going to just sort of jump right into it. For those of you that have not either designed or been directly involved in a site-read process, typical site-read

processes, I'll talk about phase 2/phase 3 oncology trials, some of the issues which impact site-read quality and auditability, and maybe some processes to improve the site-read quality, which I think will be necessary should we rely more on them for support of primary endpoints.

There are some methodological issues that impact the quality of the data coming from site reads, and this comes from both involvement in a survey, which I think the committee has, and prior experience for the last 20 years. The radiologist does not routinely and uniformly perform trials, specific target lesion selection, measurements, and complete the CRF. In many instances, a non-radiologist PI selects target lesions, measures, and completes the case report form, and even site coordinators.

A survey that was performed about two years ago over a large number of oncology sites, up to 20-30 percent of the site reads were performed by study coordinators, looking at the clinical reports, identifying target and non-target lesions.

These are non-physicians determining essentially response and completing the CRF at the site. Some of the other issues are that the radiologist is generally not included as a co- or sub-investigator. So, essentially, the tumor forms that are used to transcribe measurements in these site reads in this CRF are really performed -- that transcription is performed by the study coordinator.

Some of the other methodological issues, the standardized training for site readers is significantly less than for blinded readers. You may be aware that radiologists in training and, generally, in clinical practice do not use RECIST or similar criteria in their clinical practice and do not receive training on this criteria. So the training that most radiologists at sites receive is that provided by the sponsor for that trial. And that is all they're familiar with, unless they participated in prior trials perhaps.

I'm not here to say that a non-radiologist should not be evaluating these cases, but we've

talked about variability in tumor types and difficulty in measurements and measurement reproducibility. So a non-radiologist can probably reasonably accurately measure a nice, well-circumscribed colorectal lesion or a thyroid metastatic lesion. But more than complex lesions, lesions that have surrounding edema or hemorrhage, those that have calcifications, have had prior local regional therapy, adjacent arteriovenous shunting like we see in HCC, or simply just poor margins, really benefit, we believe, from a trained radiologist, and a radiologist at the site that is trained to accurately distinguish tumor from non-tumor imaging effects.

One thing about the site reads, do we think that these site reads are truly auditable and are they performed to GCP standards. Well, as we're aware, the source data from the site is not the CRF, but it's the image with the lesion measurements at the site. And I can assure you that in the vast majority of oncology trials that I've been involved with, these lesion measurements

with overlays on a per lesion basis are not saved anywhere in clinical pack systems at the sites. And so that makes auditability of those tumor measurements impossible.

The site CRFs don't generally include screen shots. The electronic CRFs that are often used at the sites include usually only the tumor measurement numbers but not screen shots of those tumor measurements. And the storage of these images, as I said, with the tumor measurements on the clinical review systems just doesn't occur routinely. So this really calls into question the auditability of using a site read in supporting this question. Monitoring occurs between the tumor worksheets by the sponsor and the CRF, not using the screen shots of the image's measurements routinely. So when all this occurs, the source data for the site reads I believe is really not fully auditable.

Just to tell you a little bit about our site-read audit experience -- and this is where we have been requested by sponsors to perform selected

audits of site reads from early phase, phase 1/phase 2, of multiple solid tumor studies and various other tumor types. Initially sponsors would ask us to review responders, then all subjects and findings. So some of the things that we found in doing scores of these reads now is that -- one of the biggest issues is because there's been a lack of standardized training of the site readers, the criteria is generally misinterpreted, so inappropriate both on response side and progression side.

We've seen image interpretation issues of benign lesions considered stable disease; resolving benign disease, like pneumonia, considered response; lesion necrosis inconsistently included or excluded in measurements; benign lesions identified commonly as non-target lesions; and new lesions often not new tumor but new benign processes, and many others. Now, it's not to say that these same interpretation issues cannot occur in an independent read process, but with greater reader training that we see in independent reads,

some of this variability can be decreased.

One thing I wanted to point out, as we're all aware, is the draft guidance for Standards for Clinical Trial Imaging Endpoints that was released last year paid significant attention to the independent review charter and the methodology of the independent review program for blinded reads. Our concern is that the draft imaging guidance and the audit process could result in essentially a double-standard here. While the imaging guidance stresses blinded reader qualification and emphasizes reader training and retraining, and performance evaluations -- something that the committee has certainly called into question on several studies in blinded reads in the last several years -- and detailed information about blinded review in the charter, there is no mention at all about site-read charters. So the guidance itself is a little bit weighted to independent reviews, but not to improving the quality of the site-read process.

So what do we need to do if we are going to

rely on site reads more to support primary endpoints like PFS? A dedicated radiologist site reader for each site. Add the radiologist to 1572 because if you've got someone that is contributing significantly to the primary endpoint data and they are not signing a 1572 form, we see this as a potential issue. These readers generally are not compensated, so compensating them as a sub-investigator we think is important. And significantly implementing standardized training and testing of the site readers, I disagree with some of the other statements that it's both difficult and expensive to standardize training of site readers. We've done it for large phase 2 and phase 3 trials, both training and testing.

Certainly one of the issues, though, is reader performance monitoring. If you have three or four blinded readers evaluating and monitoring their performance, inter- and intra-reader variability is something that can be controlled, and it's certainly something that is discussed in the guidance. But in a phase 3 trial with 120 or

150 site readers, reader performance becomes an interesting challenge at the least.

Sponsor and CRO monitoring teams we think should be trained the least in basic imaging and monitoring so they can monitor the images at the site, monitor the true source data if we're relying on the source data. And potentially incorporate images into the site CRFs to ensure that that process is an auditable process.

All of this could be done by implementing a site-read charter, again, standardizing the site read, establishing site reader credentials as the draft guidance recommends for a blinded reader. Develop processes to evaluate the reader performance throughout the trial. Question. Is reader variability over 120 or 150 readers actually measurable? And control the read environment more. We mentioned about training the study team to monitor that source data.

So this is just a bit of summary slide on some of the issues with on-site versus off-site reads. The off-site reads generally have a fully

audible source data. On-site reads source data can be difficult or fully not audible, depending on how they're currently performed.

On-site reads. Clinical data and supplementary imaging and studies are available. And for someone who has actually performed a site read, Dr. Dodd mentioned we don't know to what effect an oncologist standing over a radiologist's shoulder really has a difference. I can tell you it has a huge difference, from personal experience and from site readers that we work with, on both sides. This patient is doing quite well. We really don't think that that tumor has increased in size, or just the opposite. We think that this patient is clinically progressing. Don't you see something that would meet a radiological progression? I see smiles in the group here because I'm certainly not the only one who's experienced this.

Multiple site readers from each site, often not a radiologist or even a physician at site read. Standardized reader training and testing certainly

on the testing site is uncommon, and reader training often consists, in my experience, of distributing a PDF of the RECIST criteria to the site readers. Images interpreted using clinical terminology often. Clinical reports used to complete the CRF, which is not an accurate way, for those of you that have seen clinical radiology reports, to define target and non-target lesions. On-site reads, really an uncontrolled environment, and reader variability is just difficult to assess.

So we think that there are significant improvements that can be made in the site-read process. Historically, they've been used for patient management, not really providing the primary data for drug approval. If we are moving in that direction, then the quality of that data and making that data auditable we believe is critical. You can't substitute for central review for any significant imaging endpoint. We believe, though, while double-blinding may reduce bias, there's no accounting for the issue of variability, particularly with the site reads. Trained and

qualified radiologists, as I mentioned, are not always performing the image evaluation.

So one point that I definitely wanted to make is there's a concern, certainly by many in the industry, that this is just going to end up being a transfer of cost; that when we improve the site-read process to make it an auditable process, that we're going to just essentially see a pendulum swing. And we need to be thinking about better ways to improve what we're already doing, and that is in the independent review process. Thank you.

DR. SEKERES: Thank you very much. I'd like to invite speaker number 2 to the microphone.

MR. BUTZBACH: Thank you for this opportunity to speak. My name is Arnaud Butzbach, and I wanted to say I'm a full time employee and shareholder of MEDIAN Technologies, providing imaging services to the pharmaceutical industry and investigator sites, and also medical devices, making software for the same sites in their clinical routine practice. MEDIAN Technologies is a public company, which is also part of alliance

with Canon and Quintiles, which is a big player in this area.

I'd like to start a little bit, the issues with radiologic assessment, both at investigator site and independent radiology, independent radiologic assessments. I'll try to explain a little bit what could be the desired characteristic for radiologic assessments and provide a tentative description of a paradigm that will fulfill those requirements. I'll then describe how it comes into practice and maybe discuss a little bit the implications for the random sample IRC paradigm, which is discussed here. And then I'll provide some suggestions.

Medical imaging is key to oncology trials, and there are a number of difficulties, as was explained today. Some of the radiologic assessment issues, one is the fact that patients were treated as having no measurable disease, and it could account up to 9 percent. There is missing data, very often, up to 10 percent, missing images, and there is a very high discrepancy, 24 to 39 percent.

What is quite interesting is to see that lesion measurement is not very highly contributing to this, only 10 percent. Other contributing issues could be the way people assess -- reviewers assess the new lesions, for 30 percent selection of lesions and also assessment of non-target lesions. And there are also imaging issues.

There are consequences for the overall quality of the data. It requires multiple reading, censoring bias, larger sizing, potentially, of studies, wrong decisions made -- in some cases, for drug development or maybe approval -- and additional costs and delays, and lost revenues, of course, for pharmaceutical companies. So there are impacts on drug approvals. There's a famous reference with Avastin where there were 10 percent missing scans, and I've heard of patients not followed until PFS. So this was revoked by FDA, recently.

So there are simple and complex issues together. When I say simple, it's not simple to solve but simple to apprehend; missing data,

obvious transcription mistakes, lack of compliance to the protocol, low involvement of local radiologists. Quantitative assessment is not required in clinical routine most of the time, raising the validity of the quantitative measures and validity of imaging, the criteria for the indication and drugs. RECIST is not always a good criteria, depending on the mechanism of action. And as we heard today, validity at the patient level is not there, so we cannot use this assessment to decide on patient more reliably.

So imaging techniques and modalities are not available at all sites, and there is no reference data to evaluate the radiological assessment through the drug development process because there are changing imaging techniques over the course of the drug development.

So the first conclusion is that those issues all contribute to the problem and cannot be addressed separately and should be applied to both investigator assessment and also independent review.

So the desired characteristics for radiologic assessment will be the accuracy of quantitative assessment of PFS, making the patient data available, correctly image patients, measurable, no image lost, and precision is required through improved reliability of measures. Of course it should be able to implement criteria and imaging biomarkers for measuring PFS adequately and consistently, with indication and mechanism of action. So implementing different criteria. And it will be comparable to the drug development process, comparable between early phases and late phases.

Last, but not least, the ability to cope with the trend toward targeted therapies in oncology and personalized cancer treatment, where everybody benefits, where radiological assessment is being valued at the patient level. Of course, any single reviewer, wherever located, should be able to perform such assessment, not only investigator but also IRC.

Expected benefits are quality time and

costs. For example, IRC not being mandatory anymore, less patients lost, ultimately decreased study sizing and duration, better source of information for making informed decision -- drug development and approval -- and ultimately applicable for routine treatment. We want to present a paradigm that will fulfill those type of requirements. Typically, we suggest standardized and computer-assisted radiological assessment tools to help readers at sites and also for independent review; using criteria and study-specific structured reporting; enable and enforce longitudinal assessment; embed quality control as upstream as possible; and provide computer assistance for gaining time and accuracy. For example, longitudinal assessment.

So making sure the same lesion is followed up, or making sure a new lesion did not exist profusely, which is a very common issue; help in lesion measurements, typically not only one dimensional but volumic measurement; help in lesion detection through detection systems; and ease

repetitive, complex tasks, measuring multiple lesions, computing statistics automatically, et cetera. Such a system will enforce reviewer compliance, including through edit checks and coherency checks.

So it is very important that advanced imaging biomarkers and multiple different criteria be implemented in such a way to have more inherently reliable measures. An example given during this presentation was the volume versus the diameter of tumors, and it should reflect the effect of drug in patients.

Of course, as mentioned before, this needs to be integrated into the clinical trial workflow and available to all reviewers. Images and assessments results should be collected in the course of the radiologic assessment and centralized, which could solve the audit issue and, of course, be available at all investigator sites and IRC.

The quality of such a system is probably an issue, so we suggest considering using cleared

medical devices. The role of industry is really important to the process of standardizing radiological assessments in clinical trials. Imaging CROs could be devoted to make this happen. There are a number of specificities for each trial, and a specific expertise and knowledge of clinical trials is required. Investigator sites do need support and training.

Such a paradigm comes into practice first by doing some prospective trials, and some of them are already ongoing, where standardized imaging assessment is implemented at both investigator sites and independent review. So this will demonstrate first the feasibility of such an approach, the increased quality and accuracy, et cetera, and cost efficiency. There are a number of issues. Assistance must be available to all sites. So there are solutions: smart deployment and cloud computing. Adoption by stakeholders would be a consequence of proof-of-claims data available. And, of course, it will present a shift in the industry, and probably industry is reactive

enough to jump on the bandwagon.

Coming back to the random sample of the IRC paradigm, radiological assessment is key in oncology. There is a tremendous unexploited value of such data. Industry obviously could not afford relying on non-radiological criteria only, so test survival. The current IRC paradigm is not satisfactory. The patient treatment is always done at the investigator sites, and there is no industry driver for improving the investigator site quality of radiological assessment.

The various issues with random sample IRC audit have been extensive this past year. One thing we noted is there is only a negative possible outcome, only from the result of a study, so nothing to demonstrate the effect of the drug. It's not clear which hypothesis will be rejected by such a test and what to do if there are no conclusive results. Will we continue, extend the sample, or reject the study?

When to extend the sample to all scans? Obviously, when the analysis of the sample is not

conclusive about what to do, and then it becomes cost efficient again. What to do, again, when the sample is adversely conclusive? If the beginning of the sample says there is a bias, what do we do? Do we continue doing the full scan to confirm this or to inform this?

Our recommendation is strongly positioned in favor of standardizing the radiologic assessment at investigator sites and for independent review. Implement a transition where both assessments will be done in parallel, standardized, meaning the tools; consider ongoing evaluation of discrepancies and possibly adjudicate between investigator sites and independent reviewer; consider feedback from IRC to investigators to improve assessment quality; and consider adaptive IRC design, as also discussed before.

The first, probably is to get rid, totally, of IRC once supporting data is available. Maybe the most important is to leverage those improvements for the benefits of public health; that is, enable reliable post-market, routine

radiologic assessment, which does not exist today, and support targeted therapies and personalized medicine.

Thank you for your attention.

DR. SEKERES: Thank you. And we'd like to invite speaker number 3 up to the microphone.

MR. RAUNIG: Good afternoon. I want to thank the committee for giving me the opportunity to speak. I'd like to comment on the statistical presentations today. I think they were outstanding. There were a number of issues brought up, and I would like to talk today about some of the statistical inconsistencies that I've seen in both the briefing document and some of the conversation going on here.

For reader variability, an investigator's site typically uses a single reader. And the reason why an IRC would use multiple readers are twofold. One is to reduce the amount of bias that may be included in using a single reader in a paradigm that includes a very small number of readers, maybe up to six, maybe only three. Where

an investigator site using 120 readers, that bias would be washed out, hopefully, with trained readers, not discounting the fact that some of those might not be read by true radiologists.

The other reason for an IRC using multiple readers is to decrease the variability. So using a multiple 2+1 paradigm where the adjudicator would adjudicate any disagreement, that variability falls. And assuming that you have a standard, typical, 80 percent accuracy rate for a single reader, you would expect to get disagreement of about 30 percent, and that's about what we're seeing.

For slowly progressive diseases, you expect to get much more. And for frequent visits, you could expect to get 50 percent disagreement. So disagreement is not an indication, necessarily, of poor reads, it's an indication of many things. It's an indication of reader variability, image variability, and the like.

Just because you don't measure that variability by using a single reader at a site

doesn't mean it doesn't exist. It just means that you've washed out the bias, but the variability stays there from a single-reader concept. Each reader still has their variability, and that variability is not mitigated by the fact that you're using a majority rules or 2+1 paradigm. So the reader variability on a site read would necessarily be higher simply because you don't lose that reader variability, even for all things being equal and all readers being in the same population.

Readers at sites are not guaranteed to be radiologists. We've seen that. A matter of fact, we've seen that it's likely to be about 47 percent. In that same study that Dr. Patt showed, 47 percent were not included as the site radiologists, as the radiologists that were supposed to be included in the study.

The variabilities that would be included in any IRC read, any radiology IRC read, those variabilities, those reader performance characteristics, those reader performance problems, are endemic to readers and do not go away simply

because you go to a site. It's just that now we can't measure it. We can't measure it. We can't train against it. We don't know what the performance is. We can't mitigate it. We can't ensure that there's a learning curve, which we see in all studies, and that learning curve would lead to a more consolidated effort toward the end of the study.

As far as the burden on the IRC goes, the results shown here today are simply statements that we assume that there's a burden because we go to a pharmaceutical industry, and we say what are the problems you're going to have with the IRC. And the answer is, "Well, the IRC has a burdensome process."

Of course that's the answer because if that's the question you ask, that's the answer you're going to get. So the results here shown today don't indicate that an IRC is more burdensome than a site or an investigator site analysis. What it does show is that clinical trials are burdensome, and we all know that. That's not a

surprise.

As far as the audit is concerned, a lot of the issues brought on audits were brought up very well by members of the ODAC and the statisticians here, including Dr. Dodd and Dr. Amit. So I won't go over them any more.

If the IRC -- a simple back-of-the-envelope calculation -- and I know what -- actually, I do apologize. I meant to -- I'm a vice president of informatics at ICON Medical Imaging, and I own stock at Pfizer. A simple back-of-the-envelope calculation on what our burden is to the clinical trial is about 5 percent, and that would include absolutely no cost to the study team by using investigator site. It's about a million dollars, and that's about it, or $2 million.

If you multiply that by 5, it's $5 million. For a $100 million study, that's about the burden that an independent review would be. So a burden is not shown. I'd like to see the numbers before we conclude that that's a burden to the industry. Thank you.

Questions to the Committee and Discussion

DR. SEKERES: Thank you very much. The open public hearing portion of this meeting has now concluded, and we will no longer take comments from the audience. The committee will now turn its attention to address the task at hand, the careful consideration of the data before the committee, as well as the public comments. I'd like to ask Caleb if you could put up the first discussion point for us.

So the first item we're asked to discuss -- I'll read it again -- given the information provided on random sample-based audit strategies, the variability in radiographic measurement and logistical considerations, please discuss whether the current practice of complete-case IRC review of all patients should be replaced by a random sample-based IRC audit. And once again, some of the issues that we've heard today involve some of the potential advantages of a random sample-based IRC audit, which would be to improve the efficiency of trials by streamlining

the process, which could potentially save some money and save some time.

We've heard about the intrapatient variability in reads between a local investigator, be it a radiologist or oncologist or site personnel, and an independent body, an independent radiologic review, which, emotionally at first blush, I think causes us all some pause. But we've seen that it is systematic and systematically approximately the same percentage. And when we look at this on a population level, we see that it really doesn't affect the interpretation of trial results, which is really the purpose of why we're all here.

We've heard about a couple of potentially viable methods for introducing an auditing strategy, as well as the FDA's application of those methods to past studies and the results of that. And we've heard about just some basic core issues in interpreting CT scans, which, at the very least, could introduce some random misclassification bias.

So again, just as before, I'd really like to

hear from as many people as possible on this committee. One of our roles here is to talk about this publicly, and in doing so, give advice to the FDA on how to handle this. And I'll ask you to just signal to me or to Caleb, and we'll write your name down and go in order. We'll start with Dr. Wilson.

DR. WILSON: Thank you. I think that this whole -- I actually want to laud FDA on looking at this because I think that they've said, and I think perhaps we've seen, that there was a presumption that there would be bias in investigator reads. I think that bias is somewhat mitigated by randomized studies, but we all recognize that on control arms, particularly with crossovers, there may be a tendency to call progression early.

I think the important lesson here is that when the analyses are done, there appears to be a very good concordance between the investigator and the independent review. And I think that the more we can do to streamline trials, I think the better and the quicker we can get these trials done. I

think that it is -- I would agree that it is important to have an audit process ongoing. I think that's key. I think that if an audit process was simply removed, there may be drift in terms of the reliability. And certainly I think all of us would feel a little bit less comfortable.

But I think we heard two different methods of how random audits could be done and determine whether or not full audits should follow on. I think that's a technical issue. But I personally think that we've seen some very credible data here, trying to move to random audits in randomized studies and perhaps excluding double-blind trials. But even then, I would say perhaps you'd want to do it for several just to kind of validate even further. But I personally feel very comfortable that the notion that this is a reasonable strategy to pursue has been shown.

DR. SEKERES: Great. Thank you, Dr. Wilson. We're going to Phone A Friend now and ask Dr. Harrington to weigh in.

DR. HARRINGTON: Thank you. I largely agree

with Dr. Wilson. I think the presence of the audit mechanism has a lot of beneficial effect, not only on just making sure that people are more careful, but I think, as we heard from the public speakers, there's a technology transfer that's going on there.

It's difficult for a statistician to disagree with a proposal for doing something with random sampling, especially as in the NIH case where the random sampling will converge to the full independent review committee if there's a large enough discrepancy. I think, though, in my view, while it's terrific to look at these issues, I don't think quite ready yet -- certainly I'm not quite ready yet to say that there's a particular approach that dominates here because there are lots of things that yet need to be resolved, one of which I think in the NIH proposal, it converges to the independent review committee's view. And I think, as Dr. Pazdur said, there may not be a truth here. And as Raji said, it's important to understand the differences in the discrepancy.

I'll just close with something I think was an important subtext on what we heard from the public speakers and something I raised earlier, that if there is a truth, it's the way treatments will be administered in the clinic once approved. And I would like to urge the FDA or others to try to learn more about the variability off clinical trial and how different it might be from what's happening on trial so that we understand, when we estimate a hazard ratio in a clinical trial, whether that really is conveyed and moves to public use. Thank you.

That's it for me.

DR. SEKERES: Dr. Harrington, did you have anything else to add?

DR. HARRINGTON: I didn't. No, I didn't. I said that was all for me. I'm not sure if that part came through.

DR. SEKERES: Okay. Thank you.

Dr. Eckhardt?

DR. ECKHARDT: So, yes, I completely agree that this is really headed in the right direction,

and I think that some of the effort that has been spent in conducting the central reviews now really needs to go toward the sites. I totally agree with some of the public comment in that perhaps some of that hasn't been as standardized because there has been central review. And I think more effort does need to be put into standardizing the sites, the image analysis, and qualification of the people making the site reads.

So I think that there is a way to essentially elevate the quality of those reads such that this independent -- or rather the site auditing method -- you know, I think that a lot of methodology needs to be examined with regard to the types of audits that could occur. I wouldn't totally throw out the idea of an adaptive type of audit process, where essentially you'd be able to get real-time data as the trial is being conducted and maybe expand or contract the number of audits that are required. And I think certainly there's a lot of adaptive trial design and Bayesian methods that may facilitate that.

I do think that some of this is going to be restricted based upon the types of disease under assessment. And I'm not sure whether or not that would really mandate all independent review, but it may be that you would increase the audit rate or something like that. Certainly, there are -- we've talked about prostate cancer studies. There are many studies where the PFS is a difficult endpoint radiologically.

I think the other component that could be examined is the extent of blinding of the study because, again, sometimes that definitely is going to enter into a bias, and sometimes we're stuck with trials that are less easily blinded than others. So I think this is a real step in the right direction, and I really, again, applaud the FDA and others for presenting what I think is really credible argument to support this kind of process.

DR. SEKERES: Great. Thank you.

Dr. Buzdar?

DR. BUZDAR: I think this is a very unique

effort to maybe bring this review process on how we assess the responses in certain tumor types, i.e., the time to progression. I think the issue, which public speakers raised, is very fantastic, because the thing is that a lot of things we standardize, but here the X-rays, everybody has their own machines. When we get a drug which is under evaluation, if it has any potential to cause any cardiac arrhythmias, EKG machines are provided at each site so that it is of same company, same type. Over here, you have -- even within single institutions, there may be 20 CT scans, and the images are done -- there are differences in the quality, differences in the images, and it is difficult to compare.

I think it is not that we need to drop some layers and make it more murky, but I think we need to raise the bar and make it much more robust so that when we look at the data -- I think the point which was being raised, that maybe those selected images at initial, when the patient is being entered into the study, if it is a lung lesion, or

if it is a liver lesion, it should be photographed so that somebody can see it, instead of having 4 x 5 centimeters lesion in the liver. The next person cannot see which section they were looking at. Those kind of cuts should be visible and they should be part of the source document, which should be visible to the regulatory agencies and the reviewers. That way you don't need multiple reviewers, but same data -- FDA can look at it. Anybody can look at it. The sub-investigators can look at it, and I think that will make the process far more better than just dropping -- making the process even more liberal, I think that will make the process I think less user friendly and make it much more murky.

DR. SEKERES: Nice points. Thank you.

Dr. Liebmann?

DR. LIEBMANN: I think I'm going to echo a lot of the previous speakers, that I'm going to come down and say that, yes, the current practice of complete-case IRC review can be replaced by random sample-based IRC audit, which I think most

of the previous speakers have said. I am reassured by what's been presented here, that the current practice of investigator review is remarkably valid and has been validated by the full IRC review sort of gone on in the past.

Having said that, I agree with the previous comment that to completely abandon audit would very likely open up clinical trials to the problem of completely eliminating IRS audit of tax returns. And although I think that there's going to be a lot of technical components of how to implement an audit process -- and I realize that that's something that's going to be reserved for another time -- I want to clarify one point that came up from the FDA, which is that currently there is a push to have study sponsors determine beforehand whether or not the results from an investigator analysis or from a central review is going to be the definitive result.

So how would that play into an audit process? Would it then be expected that if there is a full, 100 percent audit, that would be the

definitive result?

DR. PAZDUR: The investigator audit would be the primary endpoint of the trial. The radiologic review, the IRC, is an audit, that it's a method to determine whether a bias is present or not.

DR. SEKERES: Great. Thank you.

Next, Dr. Menefee.

DR. MENEFEE: So I agree with a lot of the sentiments that have already been mentioned in that the process of evaluation should be streamlined. However, I guess I'm a little bit more on the conservative side of the spectrum. I am still concerned that the lack of complete-case based reviews are going to have an impact on bias, kind of almost like -- since we're doing a lot of government analogies -- looking at the TSA in airports. We knew there was a problem before we implemented the TSA, and we haven't had as much security issues. No one would think about getting rid of the TSA because we know that it's effective. And I think we should perhaps look more carefully before we consider getting rid of IRC complete-case

base. And I certainly understand the need from a cost perspective and, again, for making the studies more efficient.

One thing that was mentioned earlier, and I just would like to give more clarity, all the cases that were analyzed here were done on studies that were associated with complete-case IRC reviews. Certainly, there were studies, phase 3 studies, that were done prior to the IRC complete-case reviews being implemented. Perhaps an analysis of some of those studies might be informative to let us know if we get the same degree of conformity in terms of hazard ratios with retrospective analysis in those situations. And I don't know if any of those things have been considered previously.

DR. SRIDHARA: I think that will be a question for the sponsors because the older data we had not -- probably most of them, we had survival as the primary endpoint, and we were not putting so much emphasis on having all the data on progression and how much follow-up was there, and whether they were missing assessments and so on and so forth;

although we have included at times, in the product label, the investigator-assessed PFS information in the label. So if the sponsor has those original scans and now they're willing to go back and do an independent review is a question that I don't have an answer to.

DR. SEKERES: Okay. Thank you.

Dr. Fojo?

DR. FOJO: I agree with everybody. I think if the investigator assessment were therapeutic, we were all going to vote unanimously that it's a great drug and approve it. The only question I would have about one -- I mean, the FDA certainly is not uniform with all companies and with all submissions. This seems to make that uniform. And I would think it would be better if there was, in my opinion, some more leeway. Some you may still want to do full IRC review, some you may want to do an audit, but there might be some you feel really comfortable with in saying, no, we don't need that here.

So you're nodding your head. I'm assuming

that that's how you view it, that you can move away from that, right?

DR. PAZDUR: I think we meant this as a more general question rather than an absolute, all or none type of situation, because obviously there are issues, depending on tumor types, that one may have: size of trials, the endpoints, other corroborating evidence that might come into play here. So this is more of a general question that we're after here.

DR. FOJO: Okay.

DR. SEKERES: Thank you.

Dr. Choyke?

DR. CHOYKE: Much of what I wanted to say has been said, but I was very impressed by the strength of the correlation between independent reads and the investigator reads. And it really called on the assumption that the investigator is inherently biased. And you think about it, the investigators are really disconnected from whether the drug should be approved or not. They're probably less biased -- as a non-oncologist, I

think I could say surprisingly much less biased than I would have expected. So the idea of an audit makes a lot of sense. It reduces the burden of cost, which can be good for all of us.

The only thing I want to add is that it seems like we've been proposed two different methods of auditing, and I really like features of both of them. The NCI has a certain simplicity to it that I think is very nice, but Dr. Amit's proposal has this concept of measuring bias within the data set. And I think that's an important thing to capture if we're going to go to an audit-based system.

DR. PAZDUR: I'd like to address that. It's not an inherent bias. It's really the encroachment of -- bias has kind of a negative terminology to it. It's almost the encroachment of a uniform or a unilateral subjectivity, leading something in one direction. And that's what we're after. It's not that somebody is deliberately doing something wrong here. And I want to make this quite clear for the public. It is a creeping in of a subjectivity that

is going in one direction that we're after.

So it's not inherent, and by no means are we saying that this exists in all trials, or in most trials. But here again, when you are going to be making a major decision of licensing a product and all of the implications that means on one trial, one better have a good understanding that that is a true finding that one is really basing that approval on.

DR. SEKERES: Thank you for the clarification.

Dr. D'Agostino?

DR. D'AGOSTINO: As others have said, most of my comments have already been stated, but I'd like to sort of give my summary of it. First of all, I want to remind ourselves that, as was mentioned with the TSA example, the data we're looking at had the review done, the IRC review done. So they're better than what would happen without that having been the case. The other issue that I'm concerned about is that there's no discussion -- though it came up back and forth

here, there is no discussion about the type of tumors that are going to be measured and so forth. I think that we need to get the message across -- and I want to do it -- that we're not giving this a blanket approval or a blanket enthusiastic response.

The third item is that a lot of what seems to be happening, if this goes, it's going to be shifted to the sites. And with the standardization on the sites' part -- and it's not really clear to me that there's going to be a cost benefit -- the investigators will have to sort of do a very tight standardization so that there's credibility in what's being done, that it can be done. But that isn't necessarily what's happening now.

The other issue is that I think the auditing needs to be done -- in all cases, we can begin to come up with -- we said there are some cases where it's clearly not needed. But I feel uncomfortable saying that, yeah, there are going to be cases where it's not going to be needed. I think another issue that is important, and I mentioned earlier,

is that you can't just do a random sample. When you pick the method that you're going to look at, you have to make sure that there's broad representation. Over all the sites that are involved or represented, there are some procedures that have to be thought out in terms of how you're going about doing this. When you have a complete audit, you don't have to worry about it. But once you start saying you're going to have an incomplete audit, random sampling, then how do you do the stratification? And these items really are going to be needed.

The other point -- and it was mentioned earlier -- is that there may be some consequences of what we're saying when you shift away from this audit and things you can't do anymore, things I mentioned, tumor analysis and things of this nature. And I think that has to be thought out before one sort of plunges into saying this is a great thing to do. And then I think the procedure for the sampling and the way it's going to be done, if every company brings its own procedure and so

forth, I think there could be a lot of chaos on that. And I think there has to be come clarity.

Lastly, I'll go right back to the beginning. The word "should" is up there. I'm not seeing that the "should" is really driving me. I think it could, and I think there are good arguments for it being done. The switch from could it be done to should it be done I think has to really be thought out and carefully addressed. And I think in the end, there's good justification, but all of these points that are being raised around the table, giving and sort of buttressing our answer or supplying details for our answer, I think have to be really considered. Thank you.

DR. SEKERES: Thank you.

Dr. Logan?

DR. LOGAN: So I think it's pretty clear here that we've seen a pretty strong correlation between the investigator and the IRC hazard ratios. And the implication for that, as has been recognized around the table, is that in many cases, a full IRC analysis will not contribute substantial

new or independent information beyond the investigator assessment. That being said, the audit is very important to assess that correlation and make sure that that's the case, or similarly assess the degree of differential discordance. But I think it's also important to consider how much that's likely to impact the results. So the decision to do a full IRC versus an audit only should depend a lot on the sensitivity of the final results to what's found in the audit, especially given the reliance, as has been mentioned, oftentimes on a single, probably unblinded trial.

Dr. D'Agostino has mentioned a number of -- raised a couple of reasonable concerns. I think we should, in general, use an audit approach that is appropriately conservative. There is no benefit other than cost savings. There's no benefit in terms of determination of benefit on patients. So the real benefit here is cost savings. So that being said, things like full sample is probably appropriate when there's a modest effect on investigator progression-free

survival. As an example, the Dodd approach generally tends to default to that.

We should be careful about specification of the clinically irrelevant factor. Often the magnitude that we look at is viewed in the context of a benefit-risk assessment. So a value of .9 that's been thrown around here may not be appropriate when there's considerable toxicity and things to think about; as has been mentioned, the consideration of the appropriate disease. So the default approach should be appropriately conservative here, I think. Also, has been mentioned, the threat of a full review needs to be maintained. The threat of the audit needs to be maintained. Simply defaulting or switching at some point to no auditing is not really a good idea.

As Dr. Harrington mentioned, I think it's difficult to decide on an actual choice of auditing strategy at this point. I think both of them have their merits and further investigation is warranted.

DR. SEKERES: Thank you, Dr. Logan.

Dr. Armstrong?

DR. ARMSTRONG: I just want to speak up on behalf of the investigators who have to, in real time, sometimes with very little time, make these decisions. And since I'm one of those people who has to do that, it's reassuring that, whether there's bias -- and you're right, Dr. Pazdur. It's sort of got a negative connotation. But, ultimately, the independent radiographic reviewer, who essentially has to do this in a vacuum, which I don't envy at all, and those of us who have to do it in the Gestalt of everything that's happening with the patient and their symptoms, et cetera, that we ultimately end up with a pretty even decision, in that if we err one way one time, we're erring another way another time, if it's error. And I think the data from that is fairly consistent and reassuring, and I think that's a good thing.

I guess I would argue a little bit with Dr. Logan that there is more to having to do this than just the cost. There's the staff time in terms of getting scans sent in. The IRBs are very

involved in privacy protection when the data for a patient gets sent out. There is a lot to it. There are penalties if you don't get it in time. There are extra hassles. If you have patients that have to have outside scans, then they have to be ready, your institution, then they have to be sent. So it's not just the cost. So there's a lot to it.

But, overall, I think this is actually a good move forward. I would just reiterate what people have said, is that there needs to be very clearly defined criteria for what the audit is going to entail. And, again, I think Dr. Eckhardt brought up sort of an adaptive design, and trying to anticipate potential problems and changes in the auditing that might happen with that I think is a good idea. And the one other thing I would say is that we've been talking today really pretty much exclusively about CT scans, and I think we have to have a lot of caution about extrapolating this to other kinds of imaging.

DR. SEKERES: Great points. Thank you.

Dr. Wilson?

DR. WILSON: So I've heard from a number of panelists, and we heard from the open public hearing, there seems to be, from some, this kind of migration of monies and efforts from the independent audits to training and monitoring and all this stuff for the individual sites. And I just want to say that, to me, we have to be very, very careful. What the studies we've seen have said is that, in fact, the investigator reads are clinically accurate. And to go in and to try to fine tune, require training, more paperwork, more uniformity, not only across sites in the United States, but in Europe, the Far East, et cetera, all of our data would say that none of that will make a wit of difference, and it will encumber these trials I believe even greater than an independent review.

To me, I am much more interested in accurate designs that are unencumbered and are brought forward more quickly. And I just think that we have to be very careful that we simply don't transfer one procedure over to then requiring all

kinds of, it seems, unnecessary paperwork on investigators who are already reading these and already encumbered with a lot. In fact, I think one of the things that was said is at the end of the day, when these drugs do go out into the community, it is the very doctors that are going to be determining when to start them, when to stop them, et cetera. And that's the real world. And if the independent review committee showed that the progression-free survival is being accurately read, in terms of the hazard ratios, by the independent reviews, I think that we just shouldn't be transferring encumbrances from one group to another.

DR. PAZDUR: Remember, what we're after is -- going back to the central issue -- the presence or absence of bias. We're not after some ultimate, absolute truth here of what is the true value because that probably doesn't ever exist.

DR. WILSON: Yes. I mean, that's the point I am making, that we really are -- we're looking at whether or not there's accuracy in terms of

determining differences between two arms.

DR. SEKERES: Dr. Eckhardt, did you want to comment?

DR. ECKHARDT: Yes, just a quick comment. I totally agree. I think the issue, though, right now, I can say for several sites, is it would be difficult for them to be audited; because I think some of those issues were raised about image capturing. So I do think that there are components to this that will require a different level of what we're doing now.

DR. SEKERES: And, Dr. Wilson, a response.

DR. WILSON: Maybe I don't understand this audit, but why would a random audit not be the same? As a regular IRC audit, you would simply send the scans in to a central area. Is there a suggestion that the nature of how these audits would be done is different? It seems to me, they'd be done the same way. So I don't see how there would be any difference, then. If you need a full audit, the sites that are involved have to be able to do them, so they should be able to do a random

one.

DR. SEKERES: Okay.

Dr. Fingert?

DR. FINGERT: My question's been answered. Thank you.

DR. SEKERES: Ms. Mayer?

MS. MAYER: So I guess I have to begin by saying that although I understand this is not the topic of our discussion, from a patient and an advocate perspective, the degree of measurement variability does not increase my confidence in progression-free survival as an endpoint in the absence of valid patient-reported outcomes and quality-of-life measures. Having said that, I am as persuaded as a lay person can be by the data that's been presented by the FDA about the consistency, even though they don't, obviously, apply to the individual patient level. The bottom line really does seem to be that blinded, independent review does not, first of all, improve this measurement of variability, but more importantly that bias is really not the concern that it was thought to be;

although, I find it interesting that the threat of review seems to be perceived by a number of members to be an important way of controlling that bias, that doesn't exist.

So my remaining concern really is with the -- I'm confused, as Drs. Eckhardt and Wilson. Now I'm really confused about whether we're talking about on-site auditing or central auditing that's done in a random sample rather than all the patients. If it's on-site auditing we're talking about, I think we've heard enough, particularly in the open public hearing, about concerns about the quality and training of the reviewers. With very large trials, with hundreds of sites, with perhaps an individual site only having a few patients on the trial, I wonder if it is an on-site audit that's being undertaken; are there really multiple people who are qualified to do such an audit, a blinded audit. Are we always talking about just one person?

It's unclear to me exactly how this is going to work. I understood the different models, but

they were models, and I'm really concerned about the practicalities of implementing this. It seems like it could be quite burdensome, especially when you get down to the level of stratifying, so you'd be representing every site. So I'd love to hear some more that would give me some confidence.

DR. SEKERES: Would FDA like to respond?

DR. SRIDHARA: So this is a central audit that we are talking about, so there is no site audit. So the scans will be sent to a central place, and a random sample will be picked. It could be patients from different sites. And what Dr. D'Agostino was suggesting was in that random sample, we have to make sure that we are not picking only patients from sites which are accruing more but have an equal representation from sites which are not accruing as much, as well, so that we have a fair sample of all of them.

MS. MAYER: Okay. Thank you.

DR. SEKERES: Dr. D'Agostino, would you like to respond?

DR. D'AGOSTINO: I was not suggesting -- I'm

sure you realize it -- that there be a site audit. It's just the sites are represented. If you have these major sites that are producing 50, 60, 70 percent of the subjects, you do a random sample, they're going to be overrepresented. And the smaller sites -- for us to sit and think that there isn't variation amongst sites is insane. There will be variation. So we have to make sure when we say we're giving blessing to a sampling procedure, that all the sites have some representation, not that every single site, somebody's pulled from that, and certainly not that we'd go and visit the sites.

DR. SEKERES: Dr. Zones.

DR. ZONES: To the extent that FDA implements the audit methodology, I'd like to see a prospective evaluation of the different audit methods to involve like -- we heard about two, but there may be more. But I'd like to see some comparison between them and use that as we forward to think about how these audits are going to proceed.

DR. SEKERES: Thank you.

Dr. Shankar?

DR. SHANKAR: So just sort of capping most of what's been discussed here, but just my two cents. I think having these two options that were offered today as potential audit options is certainly worth considering. And a prospective evaluation of those would certainly be helpful. I realize we're talking about independent reviews, whether it's for the entire data set or for a random sample. But I do want to take a moment to say that the issues brought up in the open public hearing about site reads are very important.

I think as there's a draft guidance for imaging endpoints in clinical trials, just to have some language; not necessarily mandatory components, which start to make it burdensome on the sites, but certainly a level of education about response, assessment guidelines as well as some level of engagement of the site radiologist as a sub-PI, as suggested. Things of that sort certainly should be considered so that that quality

of data that we see for all the other trials can also be improved and at least less variable, shall we say. Thank you.

DR. SEKERES: Nice point. Thank you.

Dr. Wozniak?

DR. WOZNIAK: I'll just add my two cents, too. I actually do agree with the random sample approach, especially for very large trials. And I think that the number of patients that are sampled depends of course on the size of the trial. I do think that there's also room, though, in some trials, for a complete-case review. Dr. Pazdur mentioned that we often approve drugs based on smaller trials because there aren't enough of those patients that would benefit from that drug. And I think for smaller trials, especially if the approval of the drug is dependent on these small trials, then probably a complete-case review would be appropriate in those instances.

DR. SEKERES: Thank you. Dr. Liebmann?

DR. LIEBMANN: I just wanted to agree with Dr. Wilson's comments previously. I've heard a

number of people talk about the need to standardize or somehow upgrade either radiologic assessments or radiologic facilities at investigators, particularly if we're going to move away from mandated central review. And there's nothing in the data that's been presented here that says that that's necessary. And if our recommendation to go to an audit rather than automatic central review resulted in a shifting of burden to local investigators, I think that that would be a misinterpretation of what we've seen today.

I just wanted to make one other comment. Although I'm not happy with just abandoning some oversight altogether, including audit, and I invoke the IRS as an example of where audit is necessary, I agree with Dr. Pazdur that that really has to do not at all with my concern about the sort of innocent, subjective interpretation, not bias, but it has to do with the reality that people are people. Human nature unfortunately being what it is, there have been cases of fraud, and I think that it is important to have an oversight component

to make sure that that is caught.

DR. SEKERES: Okay. We're going to take two more comments, then move on to the next issue. Dr. Fojo?

DR. FOJO: So, again, just to clarify, I mean this is mostly to unburden you, if you will, of the need to have IRC as sort of a standard process. And I would imagine that you envisioned this as something that will be a work in progress. I mean, part 2 there refers to -- I think we would probably all agree, a small trial should be --

DR. PAZDUR: It's not to unburden me, it's to unburden you.

DR. FOJO: Right. But with regard to 2, we would probably all agree that a small trial should be audited, as you were pointing out. And certainly one that had response rate as an endpoint should be audited. I don't know if you all have given thought to -- for example, in the crossover design, which has been suggested, whether or not the audit could be biased in terms of you want to look to see, those who came off earlier rather than

those who came off later. You might sharpen your focus to where you think there might be bias, right? And you might have something in the data that suggests that.

DR. SRIDHARA: So the discrepancy -- the discordance that was mentioned by the PhRMA group kind of looks into that. And since this is progression-free survival, the crossover side of it is generally not included. What we have seen is most of the events have progression and not really survival in the PFS endpoint, although it could be that as well.

DR. FOJO: But I think I was referring to trials where you had crossover, not the group as a whole. But it's okay. I understand your point.

DR. SRIDHARA: Yes. So the crossover happens at the time of progression. So for the purpose of this endpoint, it doesn't matter because they have already reached the endpoint for evaluation.

DR. FOJO: Right. Okay.

DR. SEKERES: Dr. Wilson? Okay.

So I'm going to summarize the comments that I've heard about this from everybody. And thank you so much for -- I think everybody volunteered to say something and had a lot of great insights.

People commented that investigator review is remarkably valid, which is reassuring. You talked about standardizing the quality of reads and technology at sites, the idea of adaptive audit in real time; study blinding; incorporating images into source documentation, that we shouldn't completely eliminate secondary audit or review. And I think both the IRS and the TSA analogy could apply to this; call for flexibility in using this in some instances but not all, particularly with tumor variability and with small trials.

There's a concern of shift of costs and procedures and additional requirements to investigators and sites that I think everyone uniformly agrees should not happen. Auditing should not be a random sample but should take into account variability of sites, and auditing should be conservative. Full samples should be invoked

with a modest progression-free survival advantage or small studies. We should think of this not in terms of just cost savings but also savings in personnel time and frustration; caution in extending this to other radiographic modalities; concern about central versus site audits, and that we should consider a prospective evaluation of the different auditing methods.

So, thanks everyone, and let's move on to the second discussion point. Please discuss situations where a random sample-based IRC audit may not be appropriate. Some of the things we've heard of so far have been, as we mentioned, modest progression-free survival, small studies, and people talked about some variability in tumor assessments. If people could be specific with that, that would be helpful. I'll start with Dr. Wilson.

DR. WILSON: Yes. I think you just said it. I mean, I'm not sure we can be specific about which tumor types, but I do agree with all of what you said; exactly right, small trials, certain tumor

types that are difficult to assess and responses, although that's not something we're looking at. I don't know. I mean, one thing we discussed earlier on, are there specific tumor types we could discuss and bring up. There are some examples one can come up with, but I think it's probably going to be on a case by case and setting by setting.

DR. SEKERES: Okay. Thank you.

Dr. Steensma?

DR. STEENSMA: Thanks. I got skipped over on the first one, but what I had to say equally applies to this one. It was like being at Christmas and Santa Claus gives presents to everyone else.

(Laughter)

DR. PAZDUR: Choyke also did not talk, by the way.

DR. STEENSMA: Oh, really?

DR. SEKERES: You're the only one with the lump of coal, Dr. Steensma. I apologize.

DR. STEENSMA: I probably deserve it.

What I was going to say is that I don't

think a one-size-fits-all approach to this is workable. I think that we have not seen evidence today that there's any sort of systematic problem in investigator interpretation as a result. Although audits keep people honest and they're necessary to continue for that reason, I think universal -- the complete-case audit is not necessary, except in special circumstances.

We've heard a couple suggestions about particular tumor types, and that discussion is probably too complicated to have here. But I think the agency needs the flexibility to require a complete-case IRC review and, say, a trial of POEMS syndrome, where radiographic evaluation is notoriously difficult, or carcinoid, or some of these other situations, but then allow investigators to do what they do in other sorts of settings.

I'm very sensitive as a clinical investigator to the burden that any sort of additional training or regulation that's mandated puts on the investigators, as well as the sponsor.

And to use one final analogy to the IRS and the tax code, there's this whole industry of CPAs, millions of CPAs because the tax code has gotten so complicated, that most Americans would rather pick up a rattlesnake with their bare hands than do their own taxes. And I think the danger is that with each additional level of complexity, mandatory training, whether it be required by the regulatory agents or sponsors requiring things as part of an extremely conservative CYA interpretation of the regulations, it just makes it more difficult to get the studies done that we need to do to improve patient outcomes and help patients live longer.

So that's what I think, is that, yes, there are such situations where random IRC audit may not be appropriate, but I think with input from consultants as needed, the agency can make those determinations and just needs to have the flexibility to require it in some circumstances and have a sample-based audit in others.

DR. SEKERES: Well-spoken, Dr. Steensma, and worth the wait. Thank you.

Dr. Fingert?

DR. FINGERT: Howard Fingert, industry representative. I understand Dr. Steensma's comment, that he doesn't really want to get into details about tumor type, but I would like to ask the agency if they could help me clarify where the data are behind that, if there are any now.

Are there situations where tumor type alone would require a 100 percent IRC review and preclude consideration of a partial audit? And the other side of that question is, are there features of a trial design or program that could actually support a partial audit, irrespective of tumor type? I mean, for example, we've heard about whether or not there's arm -- there could be similar safety profiles in an add-on study in both arms, so that the safety isn't unblinding the study; if there's experience with investigators and experience with trial history in that tumor type; where there's confidence in the PFS; from an industry perspective, maybe if it was an sNDA, if the drug already had approval in third line and now you're

going to second line.

I know you're not going to really address that one, but it seems to me that it does require -- the reason they asked question 2 is they want us to discuss it. So could you give me more insight as to what you mean by this tumor type? Do you really mean that just based on a tumor indication, it would require 100 percent audit, or is there going to be special circumstances, like those that are listed, or others that you can help us understand where a partial audit could be considered, irrespective of tumor type?

DR. PAZDUR: Well, let me address that. First of all, answering the easier question, if the trial is truly blinded and we are convinced that the toxicities do not unmask the assigned treatment, no. And we have stated this to companies already, that we would not demand a central review or even auditing procedures. But here again, Howard, I think we have to be very honest with ourselves. This is few and far between of drugs that are entered into clinical trials,

that they have very similar toxicity profiles to a comparator arm. Can they exist in the future with perhaps more benign types of drugs that are more targeted? That may be a possibility, but from a conceptual point of view, if the study is truly blinded, then we would not ask for a central review either, in the past or in the future. But I think that we have to feel very comfortable that that blinding is maintained here.

The issue that I think that has been discussed here are tumors that are poorly demarcated on X-rays. And I think that it doesn't really address the issue of necessarily the need for a partial review or for an audit, but how uncomfortable people feel with progression-free survival when one has a poorly measurable disease. Examples of that would be in hepatomas where you have a high degree of cirrhosis confounding the interpretation and the measurability of the disease; ovarian carcinoma, where you may have very difficult times of measuring the tumor; carcinoid tumors and vascular tumors that might be very hard

to read and you have varying levels of sophistication of radiographs.

So I'd like to get back to the central reason of why we're getting these audits. And the audits are basically to determine a bias here, and that has to be separated from whether or not progression-free survival should be the appropriate endpoint. Sometimes we're forced to accept it, even in these poorly measurable diseases, because of the natural histories of the diseases. For example, carcinoid is a very long natural history. Charles Moertel referred to this as a carcinoma in slow motion, so to speak. So it's impractical using these endpoints, and we're sometimes forced to use that.

But I think we have to go back and discuss internally whether this really is a reason to invoke a 100 percent audit because, here again, the principle here is the detection of a bias, not whether one can measure it or not, because measurement inaccuracies should be present in both arms as noise, so to speak. So there's not a clear

answer, and the answer that you're looking for, no one has.

DR. FINGERT: Just to respond that? Thank you. I think that actually was very informative. I was a bit confused because people have raised this point about ovarian. And yet when I looked at Dr. Amit's presentation, the analysis they did, they had ovarian. And the analysis the agency did also had ovarian with very high concordance rates. So ovarian's being presented to us as one where there are examples of good concordance. So to think that just, a priori, ovarian means it must be 100 percent audit -- 100 percent IRC, that's really where my question was coming from.

DR. PAZDUR: I think really what it reflects in the discussion here is the poorly demarcated tumors and whether PFS should be the endpoint or should we look at overall survival, not necessarily whether or not an audit is necessary; because remember, let's get back to the central issue, it's the bias that is an issue, not whether we can measure it or not. Here again, if you're measuring

a poorly measurable disease, it's going to be present in both arms here.

DR. FINGERT: There's one regulatory technical feature about this that we might address. And that is, you're talking about the situation where you assume PFS is the primary endpoint. But with all this discussion today, there are times that a sponsor may want to have PFS as what we call a key secondary endpoint. And the question is -- let's say OS is the primary, and it's powered for OS, and then key secondary might be PFS. In that situation, is it your vision that what we're talking about here would apply in that kind of setting as well?

DR. PAZDUR: Well, in general, if you win on overall survival, you win. Okay? And I don't think we would demand for secondary endpoints, necessarily, a radiological review, certainly not a complete radiological review. In the past, we've included these as secondary endpoints.

What is the importance of progression-free survival when you've already won on overall

survival for the company? In making marketing claims, in the patient's use of a drug, overall survival trumps all. And if you've shown that benefit of overall survival, the use of either response rates or progression-free survival as a secondary endpoint provides perhaps corroborating evidence, but whether or not a patient should use it or whether or not one could use it in marketing complaints is almost rather a moot point. Right?

DR. FINGERT: I think that there may be times that a sponsor may want to include the PFS in the label.

DR. PAZDUR: If they want to include it, we'd be happy to discuss it with them, but I think it's relatively marginal, the benefit, if you've already won on overall survival. I wouldn't demand a review of it.

DR. MURGO: But often you don't have that information, those results, before you plan your study.

DR. PAZDUR: But you would specify the primary endpoint.

DR. MURGO: Right.

Adjournment

DR. SEKERES: Thank you, Dr. Pazdur, for entertaining so many questions.

If there are no other comments, I think we're going to bring this to a close. Thank you, everybody, for the time and energy you devoted to this topic.

(Whereupon, at 2:58 p.m., the meeting was adjourned.)