something, comparable to a radiologic physics center, or something like this?

DR. GIGER: Yes, it would be an independent center which would preserve the integrity of the database, and its output would be valued by the FDA and the community.

DR. BOURLAND: And is the expertise available there, disease-specific, modality-specific, CAD-specific, I guess, site-specific?

DR. GIGER: Well, such an institute could range in what technologies it is assessing. Specifically for CAD, it could start with the most looked at modalities, which would be mammography, lung CT, and colon.

CHAIRMAN GLASSMAN: Dr. Dodd?

DR. DODD: I have a question for Dr. Nishikawa. With regard to the reasonable FDA endpoint threshold of 3, surely that depends on the underlying shape of the ROC curves, right?

DR. NISHIKAWA: Right.

DR. DODD: Are you advocating that as a universal threshold, or would you advocate that depends on what the baseline ROC curve looks like?

DR. NISHIKAWA: I'm throwing that number out as a possibility. I'm not saying that's what the number should be. I think you -- we need to investigate this more, because I don't know what the shape of that curve is exactly. I modeled it after the DMIST data, because that's the best data we have.

And I assumed sort of the shapes were the same, which may not be true with reading. What I'm trying to say is that it's consistent with double reading. That's what -- the point I'm trying to make.

CHAIRMAN GLASSMAN: Any other questions? If not, thank you to all of our public speakers. Okay. We will now continue with the panel's general discussion of mammography CAD devices, after which they will

focus their deliberations on specific FDA questions. Following that, we will break for lunch.

I would now like to remind the public observers of the meeting that while this portion of the meeting is open to public observation, public attendees may not participate, unless specifically requested to do so by the Chair.

The Panel may ask FDA staff questions at any time. We will now move on to the general discussion portion of the deliberations. At this point, I would like to make a couple of comments, and then open it to everyone else on the panel for discussion of general thoughts about CAD, rather than organ-specific thoughts.

The first point I would like to make is the difference between CADe and CADx, in my mind. We've got this artificial distinction based on the fact that we have two different names for two different things, that

is CAD detection and CAD diagnosis.

And really, in fact, they are a continuum. The perfect CAD system would be able to diagnose all of a given disease in the case of breast cancer, colon cancer, lung cancer, and never have a false positive. Now, that would be great. It's not likely to happen any time soon, but I certainly would welcome a product like that.

But to go to the other end which is the detection only end, to make a computerized value judgment about detecting the lesion means that, in part, you have made a diagnostic decision as well, that there are certain things that the computer sees in the image that the computer ignores. To do that, that is a diagnostic judgment at a very low level.

So it seems to me, that with this continuum, the critical issue for whether something gets approved or not, is what is the use that is specified by the manufacturer. If

the manufacturer specifies a detection only role, then the bar, the level of proof, the level of efficacy will probably be different than what would be needed if someone came and said I can diagnose lung cancer.

So, I think that's something we need to keep in mind when we talk about CADe and CADx. And with that, let me open it up to everyone else on the panel. Don't forget to push your button to speak and then push it again when you are finished. General comments. Dr. Garra?

DR. GARRA: Yes, I would just like to make one comment about CADx. And that is, that remember it's a broader category, in that it includes not only diagnosing cancer, but various other kinds of lesions. So theoretically, a CADx system would not only give you the probability of cancer, but the probability of it being an inflammatory polyp or a granuloma. So you would have a complete differential, theoretically.

So it can be a broader category or it may not be. So, if you think about it in terms of one diagnosis, it's a continuum, but it could be a much broader thing. Thanks.

CHAIRMAN GLASSMAN: Other comments?

DR. TOURASSI: Also another distinction between CADe and CADx with respect to mammography is something that was raised by one of the speakers, Dr. Hasegawa. What is the management decision to be made in the end? For the CADe product, it's simply its suspicious enough to be recalled. But the envision for the CADx products is, is this malignant to go to biopsy or benign enough to be recall for short-term follow-up?

So there is a lot of interpretation in terms of how a CADx system, a diagnostic tool will be used. And this needs to be specified by the sponsor in the development.

CHAIRMAN GLASSMAN: Dr. D'Orsi?

DR. D'ORSI: Again, I would like to stress the difference between CADe and CADx.

There have been many studies that show that the correct way to deal with screening for breast cancer is to read those off-line in batch mode. And then recall areas that are suspicious, throwing out a bigger net, obviously, than you would just to get pure cancer, and then work these up.

So I think that separation is critical. And I kept on hearing a wash in and out between one and the other. And I really think they have to be kept separate. You really should not make an estimation biopsy from a screening exam, at least not in at least 95 or 98 percent.

So I worry about a mixture of CADe and CADx at the screening level, because I'm worried about reader bias. If a lesion is marked and then a number comes up, that number may be correct or incorrect and it's going to lead to more bias for that reader at a screening level to dismiss or accept that finding.

The second thing is that these CD, these CADx devices require a very large number of cases in order to work properly. And they need cases at various levels of difficulty. So that has to be something that the FDA should focus on is the number of cases that went into the training algorithm to give out these numbers.

DR. TOURASSI: Yes, actually, I would like to -- that was a very important point that was raised. That when these devices are developed, there needs to be a clear explanation to the target screening mammograms or diagnostic mammograms, absolutely.

CHAIRMAN GLASSMAN: I would like to go around the table to give everyone a specific chance to speak to this. So I'm going to start with Dr. Mittal on my left with any comments, general, about CAD, and we'll move around the table until it comes back to me.

DR. MITTAL: While my impression has been, from the discussion, that CADx or CADe has some important implications in diagnosis and in certain situations in the management of -- since I deal with malignant tumors, obviously, it's still a research tool and it needs to be evaluated further.

I deal with a lot of cancer patients that receive radiation treatment after lumpectomy and it's kind of interesting to -- and I'll sort of say comments from other Panel members is that the test data includes those patients that have architectural distortion due to radiation plus surgical resection in these patients.

DR. ZISKIN: I have no comment, at this time.

DR. WONG: I think one of the issues that I think are probably present for both the CTC, which I'm interested in, and any other CAD system, as we mentioned, would be to develop some form of regulatory process where

data can be put in in an unbiased fashion in which the various development agencies that are developing this CAD could be tested in a very unbiased way.

And I think that as we develop a larger database, we will be able to actually begin to test more in a more sophisticated fashion as to whether the CAD devices are really answering the questions that we want to answer.

DR. ABBEY: Let's see, I think that we're struggling with the issue of how to -- how we should validate these approaches, whether we should be going to full reader studies and then how to interpret the results. And I think that's -- I'm still coming to grips with an opinion on that, I suppose, and I'm looking forward to hearing more throughout the day on that.

But it does strike me that, if we are to use reader performance results, one of the things that we are implicitly getting at

is some sort of a utility of the decision. And I think that it might be helpful to incorporate that in. I think Dr. Nishikawa was implicitly doing that with his slope criterion, and I think that's formally equivalent to a utility.

DR. GARRA: I have two things that I wanted to bring up. One, I just want to know what the underlying -- this is a question more than a comment. A lot of the speakers from the public and companies were addressing the issue that, please, FDA do not change the rules and make things more burdensome for us. That was the message I was clearly getting.

And I'm just wondering from the FDA's standpoint, was that the reason for calling this meeting to possibly make things more stringent over the previous criteria? There may be a hidden agenda that everybody knows about, except me, but I'm curious about that.

And the second item was I really am

very strongly -- this is a perfect situation for a generalized uniform, carefully laundered database of high quality images performed with state of the art technology that is publicly funded. And I want to strongly support the NIH in its efforts to produce this database and also encourage the FDA to see if they can devote some funding to helping the NIH get this up and running.

I think this is perfect for that in all areas, and that's something that we're going to have to deal with sooner -- better sooner than later. Anybody have an answer to the first one, I would be glad to hear it.

CHAIRMAN GLASSMAN: Ms. Brogdon, you have a comment?

MS. BROGDON: I can respond to the first question. There is no hidden agenda here. It's not our purpose to increase data requirements for companies. This field has developed since the first CADs were approved. There have been a lot of publications, a lot

of comments from CAD users, and we are looking for what is the current state of the art.

We are looking for your recommendations on what is reasonable evidence for insuring safety and effectiveness for the devices. And as we will point out time and again, it's our obligation to be the least burdensome in our expectations of the data that needs to come to us.

DR. GARRA: Thank you very much. And I didn't mean to imply that the FDA had a secret agenda, but certainly the companies interpreted the calling of this meeting as a potential for increased requirements. And I wanted that to be stated explicitly.

DR. WATT: As a practicing radiologist, as Dr. Newstead mentioned, we are being beset by vast data banks of images and to have a CAD device to assist in detection is important. However, the use -- there has to be a distinction, and we don't want to be burdensome. The differences between CADe and

CADx, and practicing radiologists may view CADe as CADx, and that is a very dangerous thing that we have, I think, in this Panel to come to terms with to make certain that CADe is not viewed as CADx.

DR. SWERDLOW: As a former reader of mammography and CAD in busier and slower volume settings, I have a similar sort of crossover concern, not between CADe and CADx, but between acting as a second reader and acting as a concurrent reader. It is very easy, I think, for a solo practitioner out in the country somewhere to get the wrong idea about how it is supposed to be used and use it as a concurrent reader. And that poses, I think, a significant risk and danger to patients, more than any other sort of confusion.

My other curious concern would be I just can't wait until the first malpractice lawyer gets a hold of one of these.

CHAIRMAN GLASSMAN: I can tell you

it's already happened. Mostly, interestingly enough, for the defense, rather than for the plaintiffs. There have been local malpractice cases where CAD results were allowed as testimony for the defense where CAD was negative. I don't know if it changed the jury's mind, but it has already happened.

DR. BERRY: So my comment is about CADe. I want to point out that the ROC, which is, of course, a wonderful device for assessing the process is not perfectly relevant from the clinical setting. The clinical setting, there is a particular algorithm cut-point and decisions are dichotomous. And so one had ought to focus on specific points on the ROC curve.

And it seems to me that it is essential that you -- that companies show that they have improved sensitivity, which to me means statistical significance or Bayesian probability that the sensitivity is improved. This is a very low hurdle.

I appreciate the comments about the large sample size, but with enrichment, that could be improved. And in addition to showing sensitivity improvement, that you show little, if any, decrement in specificity.

And the other comment I have is to follow-on with Dr. Tourassi's point about recall rate. Recall rate is not the right thing to look at. It's what happens after you recall. Does the patient have to go to biopsy? What's the clinical consequence? It's not simply recalling.

In CDRH, one doesn't do as in drugs ask is there clinical benefit. In the drug world, you would have to show that CADs actually decrease mortality, and that would be far from -- that would be most burdensome. And I don't think it's appropriate in this setting, but some focus on what happens to the patient and the trade-off between the two.

DR. TOURASSI: Well, in my opinion, there is plenty of evidence in the literature

that CAD technology has an important role to play for medical image interpretation. There is also plenty of evidence that assessing the clinical significance of this technology is going to be extremely difficult due to the characteristics of certain diseases, like a low prevalence of cancer, as well as the very diverse behavioral patterns of the end users, which is something that we cannot easily control.

Therefore, the two issues that have been raised during the presentation of the people from FDA that we do need some form of standardization when we report results is critical. And we need to create some form of standardization as well as emphasize the training of the end users is an important part of this technology before it is deployed and fully applied in the clinical practice.

MS. FINKEN: As a Consumer Advocate, I listened to this with a slightly different mindset. I'm concerned for the

patients knowing many who have had various kinds of mammograms, CAD and non-CAD. The CAD people more likely to have the call backs and the panic that sets in and the fact that some times patients won't go back for the call back.

I think there needs to be, along with the standards, also a defined, hopefully training of radiologists and so on, to be able to work with patients where there is the possibility of a call back, educating the public to this, so that patients with a call back, even though the statistics are good, that they have not got a problem, but to not turn them away for fear that they will get that awful answer that they have cancer.

That affects too many people, and what we are trying to do is catch those early cases, which are "curable." And that's a wonderful way to approach the recall concept, I would suppose. But I think we need to keep the patients in mind as we evaluate this.

DR. SPINDELL: Hi, I've got three quick comments. And the first comment has to do more with me being a physician than being the Industry Representative. And I echo Ms. Finken's comments. But I think the first thing you have to understand is what's the risk management situation? What's the risk benefit?

And I think we really have to flesh it out. We have heard many physicians speak today, and I'm interested to hearing the expert physicians in the Panel speak. What is the risk of a false negative versus a false positive? What is the risk of -- because, obviously, these CAD products are going to increase sensitivity. We have seen multiple studies, and there is probably an increase -- a decrease specificity.

We have seen multiple studies. What is really the effect on the patient? Would a patient -- the ultimate consumer is the patient. Would the consumer rather get

called back knowing there is a much less chance of them missing a cancer or not? And I think that really has to be fleshed out by both this Panel and the FDA.

The other two quick comments I'm going to make, back to Industry Representative, is it seems we saw a couple of references to double radiology readings being the standard of care or the best practice. So would it be reasonable to consider a follow-up CAD to be equivalent to that and be effective?

And then the other thing I'm going to echo is what Dr. Garra said before about a universal dataset which would level the playing field for all CAD developers. Thanks.

DR. KIM: I guess I have two comments. The first one is that I think one of the biggest things will be training for the end user, the physician that interprets it. Because I think if it's not clear, you can run into some big problems as people individualize what CAD can or cannot do. So I think there

has to be specific sort of labeling, and indications on how the product should be used given by the sponsor and then training, so that the end users use it correctly.

And so one of the -- you know, in terms of say, the concurrent reading paradigm, one of the futures I could see or scenarios I could see is a person sort of morphing it into more of a primary read where they essentially start ignoring everything, except for the CAD marks.

And so I think, that would be a very dangerous situation. So from that standpoint, if there is a way that this Panel and the FDA can make it more difficult for very extreme situations to happen, such as leaning more toward a second reading -- reader paradigm, I think that would be beneficial.

My second comment is on that of the test database. I think it really is imperative that we have a standardized set that is large enough and that represents what

we are looking at that different companies can test against and where it is private, in that they really cannot learn from it, as we heard that can happen.

And there are a couple of unique situations happening in the colon realm that may allow for a nice database.

DR. LEITCH: I think as others have said the blending between detection and diagnosis remains an issue and how the end users interpret that. And also, how the patients interpret those findings if they are called back based on CAD.

So I think that needs to remain clear. And when applications are made, that it is very specific is this for detection? Is it for diagnosis?

The other thing which hasn't really been brought up, but if one thinks about massive screenings, the ends that screening is held to generally, as Dr. Berry mentioned, may be too stringent, that of mortality reduction.

But some of the issues of cost that goes into screening and how CAD could be valuable, either to reduce that or will it, in fact, increase the cost of screening, I think those things need to be taken into account when you look at implementation of CAD.

And one other thing which I think has been presented in some of the data that we got from the FDA about detection of lesions and over-treatment. I do think it is valuable to detect ductal carcinoma in situ. And that, you know, particularly, if you are talking dense breasts, young women, early age, that they have a longer life to live in which that could turn into something more serious.

So I think there are values to that. And if CAD is able to detect microcalcifications, that might be missed otherwise, I think that's important.

The other thing which I really haven't gotten from, I know a lot of speakers have tried to address this, is with respect to

mammography CAD. If you -- if a lot of the studies were done and the approvals were initially done on film screen, that was digitized versus now where there is more of a move to primary digital images, which, you know, when I look at those, I know I see things I wouldn't have seen before on regular films.

And when I look back at old films, and I know even DMIST and all that, notwithstanding, I do think that calcifications are better seen because of the ease of magnifying just on a routine look at those films. And so I'm wondering, does the CAD really add as much to those digital images as to the film screen digitized? If somebody has a clear answer about that?

DR. SAHINER: I think it's very important to be able to extrapolate the clinical effect of a CAD system from the data that's submitted for FDA-approval. And I think that's probably the main thing we are

discussing here. And as we have heard from the FDA speakers and others that in an ideal world, of course, you would be able to look at the data, submit it to the FDA and see what kind of an effect it would have on clinical practice.

But because of the limitations in resources, and we have heard it might take years to complete prospective studies with these systems, to me it seems impossible to directly find out what the clinical effect is going to be.

But I think it's very important to analyze this in the context that CAD is supposed to be read, which means as a second reader in a sequential mode in the clinical setting. So I think it would be important both for the FDA and the companies to encourage such studies to look at the clinical effect and then to try to reconcile it with the submitted data to understand what is different between the two, if there are

discrepancies, so as to minimize the discrepancies as the process gets more and more refined.

DR. CARRINO: One of the advantages of going at the end is a lot of the things you wanted to say have been said already, but I'll add a couple more points. I think most of us feel that CAD can be useful, but the question is how can it be useful vis-a-vis detection versus diagnosis?

I thought this morning the FDA staff did a great job presenting and summarizing some of that information. And our focus here is scientific on safety and efficacy, not economic or medical/legal.

And I thought the people who talked in the public forum also did a good job at presenting their points, but we need to determine the regulatory process for bringing these to market.

So some of the concrete things that I came up with was what should be the unit of

analysis, which I think probably should be per lesion. And if the intended use of a device is with a reader, then the testing should be done with a reader. And then also while having large databases would be very useful, it's not clear that they exist right now. So it's going to be a challenge.

DR. ROSENBERG: Yes, I'll agree that most of the points have been made. I would concur that additional large databases would be helpful for both mammography and future CAD devices for CT. And they will be difficult to obtain, but possible.

I think Dr. Nishikawa raised an interesting point about what is truth. And I think maybe that may need more research. Is truth that there was a cancer; was there truth that a recall was appropriate? And that will give kind of different help to the radiologist, and it also makes the statistical analysis very different. So I think it's an interesting point that was made. Thank you.

DR. DODD: I have a few points I would like to make. One, I think we should steer away from prevalence-based measures, such as recall rate, particularly, when we're comparing across databases or studies for obvious reasons.

I would like to agree with Dr. Berry that sensitivity and specificity are appropriate, particularly, in mammography where there are clear lines of clinical action. However, I think we should consider the reader variability here, because as we all know, there is a huge range of variability. So we might need to consider things other than our standard binomial model when analyzing this.

ROC analysis, though, I do think is also appropriate, particularly, in early development. I think if you are able to show that you actually are operating on a separate ROC curve, then obviously that should be grounds for approval.

That said, it's hard to -- much harder to determine what a clinically significant improvement in an areas under the curve is. So sensitivity and specificity are much easier in that regard.

Also, I think we need to give some consideration to which readers are impacted. If the CAD shifts readers that are less skilled, this has a potentially large impact. You might expect CAD would have a less impact on more experienced readers.

And finally, I just want to say that the concerns -- I have concerns about comparing performance across databases, even with measures like sensitivity and specificity, because these measures will too, depend on -- or because sensitivity and specificity will depend on things like case mix, which varies by database.

DR. D'ORSI: I just want to make a couple of other points. I don't think we should generalize too much between mammo CAD,

colon CAD, and chest CAD. They both have very particular problems. Colon CAD is more geographically challenged. There is a ton of areas to look at as you snake around. Mammo is relatively confined. Mammo however, does not deal with any standard findings.

We are looking for things that are very small, and the signal/noise usually is very small as well. Chest has other confounders in it. So I think we have to be pretty open when we generalize and not generalize too much from mammo down to CAD and colon.

It was -- it works. CAD works, but I think what we have to find out is exactly -- we have to fine tune more and perhaps include some of this in labeling. It doesn't work well in certain specific areas. We need more data on that. There is information that it does not work well on amorphous calcifications. It just doesn't pick them up.

We need, and I agree fully, a very

vigorously obtained standard test set that will include stress cases, standard cases, and probably omit benign cases, because our decision is basically binary anyway. This has to be really rigorously obtained and that kind of thing will let us then more confidently fine tune when and where CAD doesn't work.

On the surface it does work. It works with the no free lunch idea. You gain some sensitivity, you drop in specificity. You can't get away from that, unless you have a brilliant exam, or unless you have something new that is going to allow you to raise sensitivity and specificity at the same time.

So I think we have to define more where it is good and where it is not good with a very standardized test in a retrospective and not a prospective manner because you can control for the biases much better in a retrospective than you can in a prospective status.

DR. LIN: I just wanted to echo

some of the concerns by some of the previous speakers concerning radiologists using CAD as a concurrent reader mainly to decrease reading time versus using it as a second reader in order to increase sensitivity. I can, although I'm not a radiologist, understand if a busy radiologist would be tempted to use it as a concurrent reader in order to reduce their reading times.

In gastroenterology, we have an analogous situation with so-called capsule endoscopy. Capsule endoscopy is a technology that we have to look for occult gastrointestinal bleeding in the small intestines. The software is able to mark certain areas as being "read," you know, a higher probability of having blood.

Although the gastroenterologist is supposed to read the entire examination, it's very time consuming to do that. It takes almost an hour to do that, and the reimbursement is very low. So anecdotally,

there have been many cases where gastroenterologists have just focused on the areas that were marked, you know, by the software and really skimmed over the rest of the recording. So that's one of my concerns.

I also wanted to make a few points about standalone testing versus reader performance testing that some of the previous speakers commented on. In my opinion, standalone testing is really just a surrogate endpoint. And it's going to be important to do our reader performance testing because that's ultimately what we need to look at.

DR. BOURLAND: I agree with these various points that have been made. I think the main issues are the development of consensus or standardized databases. These should be applied at standalone and at reader, I agree, as you just stated. And I think they would also be a great value post-installation because we have heard about the training aspects, and these would follow the views as

stated by the manufacturers.

I also think there are some quality assurance issues. Most of them are taken care of in-house by the manufacturer, but after on-site, it may be, that perhaps user training is the biggest determinate of quality post-installation.

And I think importantly, also, that the match of how the standardized database is used would match the indications of use as stated by the manufacturer.

DR. STEIER: I think we're near the end of the comments. I would like to comment that I found the speakers to be very helpful with this issue and definitely very informative. And CAD sounds like a very powerful potential tool.

I did like the discussion about the benchmark and the need for setting minimum standards, perhaps even best practices for use of this technology. I also wonder about the quality control and the variability between

the different products that might be out there. Certainly, a patient may prefer the one with the sensitivity of 20 percent versus the sensitivity of 5 percent.

I imagine most manufacturers think their products are the best, and I guess there will be some marketplace involvement with that as well. Those are my comments.

CHAIRMAN GLASSMAN: Ms. Brogdon, yes.

MS. BROGDON: Thank you. I would just like to let the Panel know that FDA is not allowed to take cost considerations into account when we review devices for clearance or for approval. So we would appreciate if the Panel can stay away from that sort of discussion. Thank you.

CHAIRMAN GLASSMAN: Thank you for that advice. I was going to try to summarize what we have just said. I don't know that I can do that. I think we've got a chance to sort of vent a little bit, not vent in a bad

way, but what we have -- react to what we have heard this morning.

I think everybody felt, in general, that CAD was important. It had an important role to play, but it may be difficult for the Agency to assess that role. There was a general consensus, I think, that some kind of standardized database for testing these products was probably going to be a good idea; that there was a continuum between CAD for detection and diagnosis. And that sensitivity/specificity in ROC were not completely satisfactory, although they are the best that we have.

I think, personally, the comment that really struck with me was that to really evaluate the effectiveness, and this may not be an FDA role, but might be some interesting academic research, would be to look at readers of different levels of ability and see whether CAD affects the bottom half in a very positive way, much more than the top half.

But with that, I think it is lunchtime. So the good news is we're going to adjourn. It is now 12:15. We will reconvene at 1:15, and we will deal with the specific FDA questions that have been posed to us.

And I would like to just remind the Panel Members that we are not supposed to talk about this morning's deliberations over lunch. They are for this public room.

And also, to remind everybody for the afternoon that we have specific questions to answer, and let's focus on those without getting far afield. Thank you very much.

(Whereupon, the meeting was recessed at 12:15 p.m. to reconvene at 1:19 p.m. this same day.)

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

1:19 p.m.

CHAIRMAN GLASSMAN: Okay. I would now like to call the meeting back to order and remind our public observers of the meeting that while this portion of the meeting is open to the public, public attendees may not participate, unless specifically requested to

do so by the Chair.

We will now continue with the Panel's deliberations on the FDA questions. Following that, we will hear an FDA presentation highlighting current issues related to colon CADs and conduct the second Open Public Hearing session to give the public an opportunity once again to direct questions to either the Panel or the FDA.

At this time, we can begin to focus our discussion on the FDA questions. Copies of the questions are in the meeting handout and on the table outside the conference room. And I stress the word focus because we need to come up with some information for the FDA.

Can we project the first question, please? This is M, for mammography, 1. Please, discuss the role of standalone performance testing in the clinical evaluation of mammography CAD devices. There are some subparts to this question.

(a) If you believe standalone testing

should be requested in the evaluation of these devices, please provide your recommendations or comments on:

(i) The merits of per lesion, per view, per breast and per patient endpoints;

(ii) Whether certain substrata, for example, mammographic finding type, size, breast composition or others should be considered in device testing and labeling; and

(iii) What marking or scoring methodology should be used for reporting findings?

(b) If you believe that there are specific situations where standalone performance testing may not be important, please, comment on what those might be.

So, this is the first question that we have to deal with. And I think I would like to ask Dr. Tourassi, Rosenberg and Sahiner to lead off with this one. I don't know which one of you would like to go first. Dr. Tourassi?

DR. TOURASSI: Okay. Regarding the general question, do we need standalone testing? I do believe that it is necessary. It is a necessary part of the evaluation device for two reasons. First of all, this is the only form of performance that we have that is truly unbiased of the end user. And we need to know how this performance operates without the end user in the radiologist envelope.

Furthermore, to go back to a point that Dr. D'Orsi mentioned earlier, we need to educate the users regarding the expected performance of this device. And in order to educate them for that, we need to know the standalone performance.

And to jump to the second subpoint, whether certain substrata? Absolutely. We need to measure the standalone performance for each one of the substrata there, particularly, the type of the mammographic lesion, since there is already overwhelming evidence that

there is a huge discrepancy between the CAD performance for calcifications versus masses.

Do you want me to go over the other points as well?

CHAIRMAN GLASSMAN: If you would, yes please.

DR. TOURASSI: Regarding the -- how to report the endpoints over the per lesion, per view, per breast and per patient, first of all, if we collect the data, I don't see any difficulty in analyzing the data, providing all four endpoints. But I do believe that the lesion and per view analysis are the most important of all.

Per lesion because that affects, in the end, patient management. But per view has gained a lot more attention lately because we know that from anecdotal evidence and from some published studies, that radiologists tend to disregard CAD marks of true lesions, lesions that are marked in only one view versus the other.

So I see that there is more clinical benefit into a CAD system that has a higher sensitivity on very much basis, and if we just report the performance per lesion, not per view or per breast, that's very important patient information that makes a difference, and clinical translation will be lost.

Regarding the marking, now I don't have clear opinions on that, but I tend to favor based on the -- what the FDA personnel presented, you know, the different overlap rules and all of that. I personally prefer the scoring using the distance of the centroids, because regarding the Computer-Aided Detection systems, the whole idea is to focus the particular -- the radiologist to particular area of the image.

And I believe that being in the proximity, based on the centroid business, that would be sufficient.

CHAIRMAN GLASSMAN: What about the last part, b? Are there any specific

situations where you think standalone performance testing may not be useful?

DR. TOURASSI: No, actually, I do believe in this standalone performance.

CHAIRMAN GLASSMAN: Okay. Who would like to go next? Go ahead.

DR. CARRINO: Okay. I just have to clarify, but not as the only -- not necessarily the only testing.

CHAIRMAN GLASSMAN: Correct.

DR. CARRINO: Okay.

CHAIRMAN GLASSMAN: Okay.

DR. ROSENBERG: Yes, I believe standalone testing is necessary. As we have had discussed, it's a two-step process for CAD. One is that the device identifies the lesion, and then there is a second step which is the user interaction in which the user recognizes a CAD marking as significant or not. And that's a separate step.

CHAIRMAN GLASSMAN: But that's the next question.

DR. ROSENBERG: And that's another question. So I think this is necessary for CAD, maybe not sufficient. In terms of per lesion, I like per lesion and per view. I think as we are looking for smaller and smaller cancers, they are frequently going to be one view, only lesions. And we need to -- but if it identifies it in both views, that gives the radiologist additional information and confirmation.

In terms of substrata, I think that's an important issue in terms of improving the devices in particular and making sure they are applicable to the particular clinical situation they are being used in. So we have that information.

I don't have a strong view of scoring methodology. I think the markings have to be close enough so that the radiologist when looking at the image and doing a workup, that they are close enough. But I don't know the difference -- different

results that would be from different methodologies.

CHAIRMAN GLASSMAN: And standalone testing not being necessary for any scenarios?

DR. ROSENBERG: I don't see that as being useful.

CHAIRMAN GLASSMAN: Okay.

DR. SAHINER: So I agree with the two previous Panel Members about the importance of standalone testing. And I just want to add one more point. It is that when companies come back for amendments to their approved products, they may be able to show only standalone performance and compare the standalone performance to the previous version to be able to get the amendment.

So I think it's in the interest of the companies also to have the standalone performance in the approval process.

The merits of per lesion, per view, and per breast, of course, usually in a clinical situation, per breast may be

important because that's really when the patient is called back. It's likely that she will get an additional mammographic view. And in that situation, even if the identified lesion is not the true cancer, that -- since there will be another view, another mammo exam, there is a good chance that it will still help, even if the -- there is a computer mark there on a breast containing a malignant lesion, but the mark is at the wrong location.

And the downside is that, if for example, one does directed ultrasound, then that ultrasound may be directed at the wrong place. So I think that's something that also makes me favor per lesion analysis. And per view is, I think, important because if the radiologist sees the same thing marked in two different views and they can easily verify that it corresponds to the same structure, then it gives them more confidence that it is a true lesion.

So certain substrata, I think it's

very important because it's important to be able to analyze the standalone performance for different types of breast composition, for example, and lesion type. We heard earlier this morning, you know, calcification versus mass, so it's very important to have that kind of strata specified in the standalone performance.

And for marking or scoring methodology, I believe that there may be many reasonable ways of marking or scoring, but the important thing for scoring is that it needs to be standardized. So if a company has a PMA today and then a 510(k) a year later, they should be using the same scoring methodology.

And I think that -- I don't know if the FDA is going to come up with a guideline document or not, but if there is one reasonable scoring methodology specified, and I don't see that there is a big downside to all the manufacturers following the same scoring methodology.

And finally, for mammography, I don't think that there are situations where standalone performance is not important. I think it is important.

DR. BERRY: I agree with what has been said about, that Dr. Rosenberg says it's necessary, but not sufficient. I would go further and say that, although it's critical to do standalone, this is preliminary. This is like you don't get to first base if you don't have standalone, or since we're not in baseball season but in political season, you have to get through the primaries before you can make it to the general election.

And so in the general election or in the confirmatory pivotal trial, standalone, I mean, you can look at it, but it's distinctly secondary. The bottom line is what does it add to the practice of the reader? And I would say that in the standalone setting, it's per lesion. In the pivotal setting, it's per patient. So I guess that's

all I had.

CHAIRMAN GLASSMAN: Any other comments about this? Brian?

DR. GARRA: Yes, I would like to clarify a little bit about per lesion, per view and per breast. I would have chosen per breast, primarily because I think it's important to be able to see it on two views, especially if you're going to be planning to biopsy or something, or try to localize the lesion.

Seeing it on one view and scoring it as a hit when you can't see it, when it scores it as negative on the other view, I don't think is terribly going to be helpful. But it's a matter of semantics whether you want to classify that as per lesion or per breast.

But as long as the rules for calling it a positive in a breast, we're seeing it on two -- having it flagged properly on two views, I think that would be a good

criterion.

CHAIRMAN GLASSMAN: Let me just comment on that, if I might, you know, as somebody who reads mammograms every day when I'm not in Gaithersburg. In the screening setting though, very often we only see something on one view. And one of the reasons for the recall is to make that determination.

So that if we -- I would -- while, you know, I agree that once we have the per lesion and per view, it's easy to come up with the per breast. I think that as the other speakers have said that per lesion and per view in the screening situation is really what we're looking for based on what happens in mammography in daily practice.

DR. STEIER: I just had a question really about Item letter b. If a product were approved and then there was some degree of change in the product, substantial or not, could that trigger another FDA review? And if so, would that trigger another standalone

process if that's what was approved? Because in that situation, there may be a question as to whether or not a complete standalone process was necessary.

CHAIRMAN GLASSMAN: Ms. Brogdon, please.

MS. BROGDON: I think we are generally assuming that a change in the algorithm would result in an application to FDA as a supplement or something to the already approved or cleared device. So our question to you is what would be your recommendations for data on these changes?

DR. STEIER: All right. That does complicate it. That's what I was afraid you were asking.

CHAIRMAN GLASSMAN: Dr. Ziskin?

DR. ZISKIN: Yes. I have two comments. One, is I think that per lesion is so critical. Without some demonstration that's just an effectiveness in detecting the lesion, I would have very little confidence in

how to interpret anything else.

The other thing I wanted to talk about is the importance of the second point, the substrata, as far as collecting the sample of the training set and evaluation. If it didn't have a good representation of various substratas, you could have a false indication of over-optimistic view of the accuracy.

For example, if you didn't have any dense breasts in your testing sample, you may be misled or vice versa, it could be you have too much as far as being difficult as opposed to the general population that you would see.

DR. D'ORSI: I just want to amplify what Dr. Ziskin and Dr. Tourassi said. I think we have to go back and fine tune where these things work and where they are mid-level working and where they don't work as well. And I think it's extremely important for substrata separation in dense breasts, amorphous calcifications, architectural distortion. And I think it should be per

lesion.

DR. CARRINO: I was wondering if that per lesion analysis would include per view, so you have a lesion that's either defined on either one or two views, and then it would include it in both.

CHAIRMAN GLASSMAN: I assume that that would be. If you collect per lesion -- I mean, per view data, it would be very simple to decide whether the mark in each view was the same lesion.

DR. CARRINO: And consider it per lesion, and you define a lesion either seen in two views or one view and then you obviate having to separate per view versus per lesion.

CHAIRMAN GLASSMAN: Okay. There was another. Yes?

DR. LEITCH: I think the issue of the substrata for the clinician, you know, when they are faced with a bunch of marks, they are going to be more impressed by those marks if they know the pattern in which the

marks are most likely to be accurate. So I think that is a really critical defining point that needs to be clear in the dataset.

CHAIRMAN GLASSMAN: Yes. I mean, I certainly would agree with that. When I -- I know when I look at a CAD image, when I look at the CAD marks, what the likelihood is that they are correct based on what they are marking. And I also know where I have to maybe spend a little more time. Dr. D'Orsi's comment about the amorphous calcifications is certainly very much in my mind when I look for calcifications.

Those are the ones, if I have a CAD case, where I look even harder than I do for the others because I know that they are not going to be seen in general on CAD. So you know, we heard this morning that CAD does -- I mean, that cancers that are missed tend to be small masses in dense breasts.

We also know, based on what we know about CAD, that that is its weak point right

now. And I think it is very important for clinicians who are looking at mammograms to be aware via the labeling process. Maybe not the analysis for approval, but certainly in the labeling process that these are strengths and weaknesses of this device or this class of devices, and you need to be aware of that.

DR. WONG: I think, you know, as we speak about regulations in a regulatory agency, when you look at this repository of data in a standalone test, I think it becomes very important that whoever handles this data maintains complete security of this data.

The data is upgraded in a fashion in which new data that is inputted into this dataset is, you know, clarified by other radiologists so that these are good pieces of data that get in because I'm afraid that eventually you can game the system. And I think that's a very important aspect of the regulation of this, of the data and developing better CAD systems.

CHAIRMAN GLASSMAN: Any other comments about question M1? Yes?

DR. SPINDELL: Just a comment, a question building on what Dr. Berry and Dr. Steier said. Dr. Berry said that standalone testing is essentially an invitation to the dance. Dr. Steier said what happens if they-- if a company changes their algorithm? So if a company changes their algorithm, in that situation, would standalone testing be sufficient to backup the changed algorithm or not?

CHAIRMAN GLASSMAN: Anyone want to tackle what is a, I'm sure, a very important question?

DR. BERRY: I think it's a review question, as they say at the FDA. And it depends if it's a minor change, you know, if you move from one point on the ROC a little bit down, it's -- I think the FDA would go along with it without the extensive, the more extensive clinical testing. But I think as a

general matter, I think somebody indicated that maybe standalone wasn't necessary. And I think the more necessary part, even in that setting, is showing that it's adjunct status. Its adjunct performance is effective.

DR. SPINDELL: I guess my -- the question is if we were -- if the product has already been approved and shown to be effective in adjunct and an algorithm change is made to perfect -- to affect the performance, the question again becomes it's already a marketed product. It's already shown to be safe and effective. Now, you're going to do a standalone to show equivalence to this new algorithm to the previous algorithm. Is that enough?

And then the other question is, again you point, it -- it's the degree of the change, then I think maybe this group could help give guidance on the amount of change or the degree of change that is necessary to say go from even no testing or standalone or

something more involved.

CHAIRMAN GLASSMAN: I would like to take the first part and pass the second part off to anybody who will grab it. And that, the first part being is standalone probably good enough for minor changes? And I think the answer is probably yes.

But the second part is really the critical one. What is a minor change and what is a major change? And I think, for me at least, not knowing anything about the algorithm that I'm being asked to look at, I would have to make a value judgment at the time. I don't think we could say, you know, with 50 percent of the formula is different; you need to start over again. But if it's 49 percent and then 49 percent again next year, you don't have to.

I think it becomes really more of a judgment call based on what the FDA knows about the algorithm as to whether a standalone test is alone good enough, or if this is a

major rewrite and then more testing would be needed. John?

DR. BOURLAND: So this brings up the issue of how you test software. Software engineering and things like this, a very interesting area of work. And there are several instances where, let me just change this one thing, and then you close the door and get out of the office. And I think we have all experienced this at work that supposedly the installation of thoroughly tested software does not function.

So the question is if you make some small change, what is the impact? So it may have zero impact. It may render something nonfunctional, at which point you would like to know that before you ship it out, things like this. So there is an overall aspect, and some of these have had medical implications, relative to computer controlled devices and things like that.

Small changes can have an impact.

Maybe it's not a diagnostic impact, maybe it is.

CHAIRMAN GLASSMAN: Dr. D'Orsi?

DR. D'ORSI: Yes, maybe we should be thinking of a very standardized test set to test what is a significant change. Theoretically, there shouldn't be much difference if you keep on passing the same algorithm with the test set, and if you make an algorithm change, you ought to be able to find what is a significant change and what isn't, knowing what -- how it did before.

CHAIRMAN GLASSMAN: True. The only concern I would have about that is the ability to teach to the test set. If you have a standardized test set and it's the same time after time after time, companies may teach the algorithm to that test set and have an inadvertent negative effect on another substrata or another type of lesion. And so it at least would probably have to be rotating test sets with pre- and post-testing of a

random test set. Something like that maybe.

DR. ROSENBERG: Yes, I was thinking along the same lines that having new test sets that are nominally similar kinds of cases would make more sense than being able to accept it by the same initial test set. Also, it depends on the interface. In other words, we're talking about differences in the software for the CAD and that may not change the reader experience, but you may end up with a different reader experience if some of the interface is changed.

CHAIRMAN GLASSMAN: Don?

DR. BERRY: So what's small? I was on one of these Panels for a diagnostic procedure where the company merely wanted to move the cut-point. They didn't change anything. They didn't change the data. They just wanted to move the cut-point from 10 down to 5.

And this was a whole Panel meeting to discuss this. And I wanted to know why the

FDA needed to bring us all to Washington to talk about moving 10 to 5. My own calculation was that if you move to 5, the prevalence was the same as the positive predictive value. And so, the -- you know, it's a worthless test.

But the Panel was split on the question. And we had this, you know, enormous discussion about what I perceived to be minutia. So I look to the FDA to make decisions about this and to not bring Panels here to talk about minutia.

CHAIRMAN GLASSMAN: Are there any other comments? Dr. Garra?

DR. GARRA: A lot of the comments that were made were ones that I was going to make. But I think you can look at the algorithm. The FDA would have to look at the algorithm and say did you do additional training on a test set to arrive at this algorithm in the case of a neural network, for instance, or did you specifically make changes

based on one of the test sets that you were previously presented?

In that case, I think a new test set would need to be presented to that algorithm. One that the company did not -- would not be aware of. In other words, what Len was saying, the rotating test set, so that people don't get wise to the questions being asked.

And that would have to be used as a criterion for passing the new algorithm. And for minor changes, I would also consider changing the set point as a minor change, and that wouldn't be something that you would need to do retraining on your original test data on. So that wouldn't fall into the category of a major algorithm change.

CHAIRMAN GLASSMAN: Do we have one more comment and then I want to summarize and move on. Yes, Dr. Tourassi?

DR. TOURASSI: I'm actually a bit confused with the discussion because you were

talking about change on the algorithm, and if that change will cause performance change to the fact that it is efficiently equivalent with the previous one.

And then we are mixing now data handling. What is training? What is testing? Let's assume that there is this standardized testing set, completely dependent sitting on the side. I think any minor change -- I'm assuming now that we have these endpoints of performance in the substrata analysis.

Any change in the algorithm that isn't reduced, as long as it maintains the performance in the substrata, then it is equivalent performance. And I assume some of those changes will show improvement in one of the substrata. Let's say better performance for masses or for architectural distortion.

Then the standalone performance, according to that criteria, is sufficient.

DR. GARRA: Can I just say a comment to that? I think that's true, except

that we don't have those test sets.

CHAIRMAN GLASSMAN: Okay.

DR. GARRA: They are not available currently. So what are you going to do in the interim until they are available?

DR. TOURASSI: Well, you --

CHAIRMAN GLASSMAN: Ms. Brogdon?

DR. TOURASSI: It is one of the points that we are going to discuss, how to deal with the --

CHAIRMAN GLASSMAN: No, no.

DR. TOURASSI: -- trained test situation.

CHAIRMAN GLASSMAN: Okay. Ms. Brogdon, you had a comment, and then I want to summarize.

MS. BROGDON: Thank you. I just wanted to respond to Dr. Berry's question or comment. We don't intentionally bring minutia to a Panel. It's very resource intensive for us to --

DR. BERRY: No, my point was that

small is up to the eye of the beholder.

MS. BROGDON: Sure. We have raised a lot of issues that we would like the Panel to discuss. Most of these are issues that have come up in our discussions with companies and in our review of these matters. So, in all of these questions that we have put to you, if you do believe that some of them are not important, we do want you to let us know about those things.

Many of these will become so-called review issues, but our intent is to draft a guidance document that can go through our guidance process, and then we have a document where we have an outline for the industry on how they can study and bring in data for their devices.

So as few -- the fewer things that simply become unpredictable review issues the better. The more we can provide guidance on, the better for the companies. Thanks.

CHAIRMAN GLASSMAN: Okay. Let me

then try to summarize our feelings about question M1. That standalone testing is necessary, but not sufficient for the approval -- or for the evaluation of a product. That it gives the unbiased look at the function of the product without the confounding effects of the end user, and that that is valuable.

The group felt that while all the per lesion, per view, per breast, and per patient endpoints were useful, I think the general sentiment was per lesion and per view were most useful in evaluating this.

I think everyone felt that the data for the substrata was really very important for analysis and for labeling, so that the end users know what it is good for and what it is not good for. And that the one comment about the centroid distance went unchallenged, I think, because we probably all agree with it as a marking.

The other very simple thing I would like to say is in terms of the marking and not

the scoring, bigger is in general better. There is nothing worse than looking at a case and then not remembering if you saw any marks on it. So I think, you know, they should stand out. It's simple, but I think it's true.

We could not come up with any instances where standalone performance testing would not be important, so we think it's important across. And we felt that changes in the algorithm should require some testing, but exactly where the line is between standalone only and more extensive testing, we didn't have a line that we could draw for you.

Is that sufficient for the first question, Ms. Brogdon?

MS. BROGDON: Yes, it is. Thank you very much.

CHAIRMAN GLASSMAN: Okay. Thank you.

DR. BERRY: Dr. Glassman, can I ask, I made a point that in the pivotal study

that standalone should not be primary. And that the primary should be in the, you know, adjunct use of the device. And this is important for somebody who is designing a trial. You know, what do you write down as the primary endpoint?

And so I would like to get some feeling from the Panel as to whether I'm, you know, a lone wolf in that regard.

CHAIRMAN GLASSMAN: I think that will come out in the next question, at least I hope so.

DR. BERRY: Okay. So the next question is going to specifically address should this --

CHAIRMAN GLASSMAN: The next question looks at reader performance.

DR. BERRY: Okay. And so, if we say there that that is primary, then that immediately suggests that the standalone is secondary?

CHAIRMAN GLASSMAN: If that's the

consensus of the group.

DR. BERRY: If that's the consensus.

CHAIRMAN GLASSMAN: Certainly.

DR. BERRY: Okay.

CHAIRMAN GLASSMAN: So let us go on to question -- I forgot. I have the next question. You don't have it yet. Okay, I'm sorry. Let's go on to Question No. 2.

Please discuss the role of reader performance testing in the clinical evaluation of mammography CAD devices, and the rest of it.

(a) If you believe reader performance testing should be considered in the evaluation of these devices, please provide your comments and recommendations on:

(i) The appropriate primary endpoints and corresponding clinically significant effect sizes, specifically to comment on the use of ROC analyses;

(ii) The merits of per lesion, per view,

per breast and/or per patient endpoints in the assessment of the endpoints for reader studies;

(iii) Whether effectiveness analysis should be conducted separately or not for cancers manifesting as masses versus microcalcifications;

(iv) Whether reading time should be assessed, and if so, how.

(b) If you believe that there are any specific situations where reader performance testing may not be necessary, please comment on what those might be.

Okay. That is the second question. Who would like to take that on first? Dr. Dodd?

DR. DODD: I will try. I'll skip No. 1 at first, because I think that's a difficult question. I want to come back, and this was a point that Dr. Berry brought up about the preference for a per patient endpoint.

I think this is a subtle issue, and I think in a prospective trial per patient endpoints are clearly important because you can begin to connect the dots between, you know, the radiologist's determination for the patient, and what happens is a sequence of events that follow after that.

When you're doing retrospective studies, it's not as clear to me that per patient endpoints should be the primary endpoint largely because we need to know if the CAD marking in this particular lesion would have caused that chain of events to follow.

And so, you know, in the absence of a prospective study, I think we need to give some consideration to per lesion endpoints in this context.

And with question 3, I'm not sure I understand the question exactly. I think we don't have to consider effectiveness separately for these various types of cancers,

but we might want to include them as co-variates in some modeling of either sensitivity/specificity or ROC analyses.

And in terms of reading time, I think that should clearly be assessed. This is an important part of something that should be included in the labeling.

Going back to the first question, which I'll just start. I hope some of you will help me out here. I do have a bias for ROC analysis, but I struggle with what is a clinically significant improvement in ROC analysis or in the area under the curve.

I have heard some numbers thrown out before like .06, but part of that depends on what the underlying area under the curve is. So I don't want to throw any numbers out because I don't want anybody to stick to a particular number. So I'm not going to throw a number out on that.

But going back to what Dr. Berry said about using reader studies as being

necessary for a pivotal trial, I do feel strongly -- I do agree with that point. And, you know because we really need to understand how the reader interacts with the CAD markings.

That said there may be situations if you do some modifications to a CAD algorithm in which you might -- in which a standalone testing might suffice.

CHAIRMAN GLASSMAN: Any other -- not any other, but who would like to go next?

DR. LEITCH: Well, I'm not going to talk about the ROC curve either. I'll leave that to the experts on that. But I do think that in the -- you know, which per lesion, per breast, it should be per patient in the clinical aspect, because the effect of these technologies added in on the population, you know, your whole screening population, how is it affected.

So I think the per patient endpoint is important in that. Not that you couldn't

use the others, but I think you do have to have the per patient.

And I do think the analyses should take into account masses versus microcalcifications, again, for the similar point in the first that physicians will be better aware of the limitations and where CAD might be most likely to help them. And so I think that is important in this.

Reading time should be evaluated because again that's part of clinical practice. How long does it take you to do this? Does it help you? Does it hurt you? Does it help the patient? Hurt you -- it helps the patient a lot? So it's worth it, you know, to have a sense of that.

And that timing should take into account not just the, you know, the reader time for the with CAD and without, but some sort of computation of what is involved in the subsequent, you know, additional workup that you do based on your findings.

CHAIRMAN GLASSMAN: Let me just ask one quick question then about reading time. As a user, certainly, I would like to know how long it takes. But if I was looking at safety and efficacy, I really don't care how long it takes. If it's very effective and I need it, I'm stuck with it. So I'm not sure that reading time is really all that important.

DR. LEITCH: Well, I think you -- again, when you are talking about screening a large population, I mean, you do have to take into account what are the efficiencies of screening a large population. So again, it's exactly your point. If you say well, I -- you spend this extra time, but in the end, this is your reward. So people know, you know, what that equation is. I think there is something to be said for that.

DR. SAHINER: I think when I read the briefing document for effectiveness, it talks about a significant portion of the target population. In a significant portion

of the target population, the use of the device were intended use and conditions of use will -- that will be a clinically significant result.

So it doesn't say anything about how efficient the radiologist is in reading those images, but whether for the patient in the end, there is a clinical -- clinically significant result or not. So, I sort of agree with you that if it improves the performance, you know, if it takes 40 seconds instead of 20 seconds to read, then you know, that's a different consideration. That's not a consideration of efficacy.

CHAIRMAN GLASSMAN: Dr. Berry?

DR. BERRY: So I think all of these things are important for one to address in building an algorithm and assessing whether or not it is ready for prime time. The issue of per lesion and the relationship to the readers, the chain of events that Dr. Dodd talked about, the masses versus

microcalcifications very important to understand whether or not the device is doing something and what it is doing.

The issue, however, of running the study, the final study it has to be simple. I mean, you can't get encumbered in saying, you know, I'm good at picking up masses. I'm not good at picking up microcalcifications. What matters is the inevitable balance of the risk and benefit. And that's a per patient analysis, that's an analysis that one has to do.

And you know, we statisticians are frequently accused of being, you know, the average. On the average, you're okay, but your head is in the icebox and your feet are in the oven or something like that. But you have -- it has to be an average. It has to be simple. And you have to show a benefit in the population.

And so focusing on these other things, although critical in development and

critical in assessing what you are doing, the primary analysis really has to be quite simple, otherwise, we statisticians won't know what to do nor will MDs and others.

CHAIRMAN GLASSMAN: Let me ask a follow-up to that, Dr. Berry. I think you bring up a very interesting point. Would it in the mind of the Panel be sufficient for the standalone testing to be used for evaluation of the substrata? But that the reader testing be a simpler examination looking at either just per patient data or per patient and the subsets and not worry about the substrata?

Is that an acceptable, not recommendation, but way potentially to do this, to keep it simple and give us yet all the data that we need?

DR. BERRY: Can I address that? I mean, despite what I said just a minute ago, I do think that there are some substrata that are critical in the pivotal phase: age, for example. Suppose for -- that we were to find

that CADs didn't add anything for women in their 40s, that their breasts tend to be dense and the -- we are -- maybe it's even negative. Maybe the false positive rate is high.

So I think that there are some really important things like age to address in looking at the utility. That already makes my simple trial more difficult, because I'm looking at subsets, and everybody knows the vagaries of doing subset analyses. But I think some amount of investigating that is important.

CHAIRMAN GLASSMAN: Yes?

DR. DODD: I too, believe in simple trials. However, I think when you get to the stage of doing retrospective trials; things cannot be as simple, especially if you are looking at enriched designs. You have to start thinking about exactly which types of cases you are going to select. Make sure that you have enough of the full range, the full spectrum of disease in the case. And at that

point, things become, by necessity, more complex.

DR. BERRY: So, Lori, would you worry about masses versus microcalcifications? Would you worry about that you picked up, I mean, your point about the chain of events being critical? Would you worry that you picked up this lesion as opposed to that lesion? And picking up that lesion led you to look more clearly, and you found the other lesion, even though it wasn't indicated on the CAD.

DR. DODD: I'll address the latter question. I think the concern I have about whether you pick up this lesion or that lesion really has to do with my inference to what the next reader is going to -- how the next reader is going to interpret that. So -- and we can't do that in the absence of a prospective trial.

So, I would want to know, and I'm not sure -- I mean, I still believe in a per

patient analysis, but I do want to have some per lesion analysis, so I know, you know, what proportion of lesions are we missing and what types of lesions are we missing.

And, if an algorithm tends to not mark the ones that we consider most important in terms of action items or -- then that's a concern.

CHAIRMAN GLASSMAN: Let me --

DR. BERRY: So the issue of masses versus microcalcifications?

CHAIRMAN GLASSMAN: Dr. Berry? Can I? Let me -- this is a two-way conversation. I want to break it up for a second. Other comments?

DR. SAHINER: So of course, we are free to discuss, you know, in what kind of substrata the data can be analyzed. But I think the specific question was only for masses versus microcalcifications. So it's not going into too much detail of, you know, what the patient age -- patient age might be

important, too.

But like the breast density or, you know, other types of pathology, but it was just whether masses versus microcalcifications because in practice we see such a huge difference between the performances of standalone CAD systems for masses and microcalcifications.

So I think in that respect because there is such a huge difference between the two and the standalone performance, I think it would be important to analyze it and let the -- again, the final purpose is to let the users know that for masses currently, the performance is worse than that for microcalcifications.

And I just wanted to clarify what I said a moment ago about the reading time. I didn't mean that it's important. I think it's very important actually from many aspects. I just don't think that it -- it may not be something that the FDA might consider when

reviewing applications.

CHAIRMAN GLASSMAN: Dr. D'Orsi?

DR. D'ORSI: Yes, I think you have to look at what this device is intended to do. It's intended to pick up findings. Let's ignore diagnostic for a minute, which we should in the detection phase. It's meant to pick up findings.

We have to know what kind of findings it does well in, what kind it's mediocre and what kind it really needs retraining in. So that, to me, you can get on a standalone, initially. You can't leave the end user out because these things are not made to use as standalone. They are made to perform with an end user.

So I think you can use the standalone to get some idea about the strata. And I agree with the timing, while it is not important for the FDA to consider because you are not considering financials and efficacy and time constraints, it is extremely

important clinically.

There aren't enough people to read these exams. So if you're going to take even 40 seconds more to read 5,000 exams in a year, that's a significant amount of time when there isn't an adequate man force to do this. So I think we have to consider what these devices are meant to do, and that's to pick up findings. Not to tell you this is malignant and that's not. That's up to the reader to decide to recall or not.

So if it falls down in a whole category of lesions, as we know it does markedly with some forms of calcifications and in architectural distortion, that's important to know. We haven't done that yet. We have done it very anecdotally. We don't know how these things work in all these areas. We have to back up and look at all of this.

CHAIRMAN GLASSMAN: We're getting way behind, so let's -- we'll look at the answers to your questions, and then I think we

have to move on.

DR. GARRA: I just wanted to comment. I agree with Carl about, you know, timing is really important, because you know what it translates into: some people not getting an exam or not getting it interpreted in a timely fashion. So if the FDA is going to consider, they have to consider timin, because it translates in a world of limited resources into a delay in diagnosis which is a critical issue for screening mammography.

We have had people that have been waiting for six, eight weeks, 12 weeks for a screening mammogram.

DR. WATT: I wanted to second Carl's statements. The other thing which I would throw in is that we need to know specifically the devices. How effective they are for masses, how effective they are for calcification, because these devices are going to be used by a variety of radiologists with differing experiences. And I think that

that's a very important thing.

And this is where the use of the ROC analyses in reader studies will be important because how an expert looks at the film with a CAD marker on it and how an inexperienced person looks at it may be two different things. An, indeed, time relating to that is going to come into the fiel, because manpower is, indeed, being decreased. And so it's going to be -- the CAD is probably in many cases going to be used as a second reader, and some may even use it as a first reader, which is incorrect, but the labeling is going to be important there.

CHAIRMAN GLASSMAN: Are there any other -- 10 second? Okay, Dr. Berry?

DR. BERRY: We're not going to approve -- the FDA is not going to approve a device to find microcalcifications. I mean, what would you put in the label? It's important to see, it's important to say this is good at that and good at the other thing,

but it's secondary.

CHAIRMAN GLASSMAN: Okay. Let me try and summarize for Ms. Brogdon what we have done with Question M2. I think Dr. Berry's comment that this should -- that reader studies should be the primary analysis is unchallenged and the opinion of the Committee.

That per patient endpoints with the reader study is very important although per lesion and per view should not be completely ignored. ROC analyses were thought to be a good thing, but we really couldn't get anybody to commit much as to how good.

DR. GARRA: Can I comment on that? I wanted to.

CHAIRMAN GLASSMAN: Let me --

DR. GARRA: You can finish.

CHAIRMAN GLASSMAN: Let me finish, okay? Whether effectiveness should be conducted for cancers with different findings, I think the answer there was clearly yes. And I think it was the sense of the Committee that

reading time is an important factor for at least knowledge for labeling.

Is there -- are there specific situations where reader performance testing may not be necessary? One comment went unchallenged. That for minor modifications that standalone retesting would probably be sufficient.

Now, there was -- Brian, you had one comment.

DR. GARRA: Well, since we didn't address that ROC issue really directly, I thought it might be a good idea to do that. One of the reasons I'm not super enamored with per lesion analysis is because it does complicate, in my opinion, the ROC analysis somewhat.

But I believe, we have some smart people here and if they can figure out a way to do it, I'm not a big fan of, in other words, a free response ROC. I don't think that is going to get at the issues that we

want to get at.

But it's important, but the point that Lori brought up earlier was very -- also important, because the shape of the ROC curve may be very unusual in certain cases, just looking at the overall area under the ROC curve is probably not going to be meaningful in itself. That you are going to have to look at the shape of the curve, and you may have to go to partial areas under the ROC curve in the areas of greatest interest to look for small differences between tests; otherwise, they are going to be swamped by the overall distortions in the curves.

CHAIRMAN GLASSMAN: Okay.

DR. GARRA: Thank you.

CHAIRMAN GLASSMAN: Ms. Brogdon, does that give you the information that you need on Question 2?

MS. BROGDON: Yes, it does. Thank you.

CHAIRMAN GLASSMAN: Okay. Question

M3. Please discuss whether there are other types of performance testing you believe should be considered in the clinical evaluation of mammography CAD devices.

Who would like to start off with that? Yes?

DR. ABBEY: I guess I just wanted to start with a question. This is where I thought the free response ROC curve might be discussed. And so I would like to hear more on why you think it won't work.

DR. GARRA: Well, it certainly works. It has been published plenty of times. But I think that the analysis of what a free response ROC curve means is a little more complicated. It's not a simple sensitivity versus one of specificity. So it's -- you have a lot of undefined quantities in a free response ROC curve. So it makes it harder to analyze.

Now, the statistics has advanced a lot since they were first published, talking

about free response ROC curves. And there may be ways around some of those, but they are not trivial.

CHAIRMAN GLASSMAN: I think that's -- if we go much deeper into that, we will be way past where we need to be here. If we could. I don't mean to stop discussion. I'm actually trying to keep it going. Any other-- let me ask a question then under M3.

And that is, is retrospective testing enough, both in reader and standalone, as opposed to prospective testing for a new application or a major redo to an algorithm? Dr. Berry?

DR. BERRY: Yes.

CHAIRMAN GLASSMAN: We're making up time here really well.

DR. DODD: Yes, I would just like to -- I agree that retrospective is sufficient. I think that, you know, some people on the marker world talk about prospective/retrospective studies, or maybe

I've got it backward retrospective/ prospective studies, you know, in the context of ongoing trials, like if the data were -- had been available from DMIST or something.

If you could collect data like that retrospectively, with the appropriate follow-up data to do some kind of retrospective analysis, that would be ideal for these retrospective studies.

CHAIRMAN GLASSMAN: Someone else?

DR. D'ORSI: I was just going to say yes, I agree that retrospective data are suitable.

DR. BERRY: Yes, I agree fully. As a matter of fact, I think it's desirable. I don't think you can do a correct prospective study, which would have to be randomized, in order to get rid of a lot of the bias, and that's totally impossible today by numbers and by people not being allowed to be assigned to one or the other.

So I think not only is it the way

to go, I think it's the best way to go.

CHAIRMAN GLASSMAN: Any other comments about M3? Is this sufficient, Ms. Brogdon?

MS. BROGDON: Yes, thank you.

CHAIRMAN GLASSMAN: Good. Wonderful. M4. The prevalence of breast cancer cases in a screening population is relatively low. Please provide comments on the practice of using an enriched dataset for the clinical evaluation testing discussed in M1, 2 and 3. And then there were a number of sub-questions.

(a) If you believe that an enriched dataset may be used for these evaluations, please discuss what you believe to be the appropriate clinical and mammographic characteristics or range of characteristics. Please consider whether the following characteristics of the screening population should be considered when designing an enriched database or stress test:

(i) Breast density: 40 to 50 percent of patients with heterogeneously dense or extremely dense breasts;

(ii) Proportion and types of masses and microcalcifications, approximately, evenly distributed with a sufficient number of additional patients with architectural distortion alone;

(iii) Size and palpability for cancers; non-palpable and a majority with size less than 1 centimeter;

(iv) Distribution of microcalcifications; small clusters of up to five microcalcifications for a third of the cases, and;

(v) Type of microcalcification clusters. According to the American College of Radiology BI-RADS descriptors, e.g., punctuate, fine linear, round, et cetera.

In addition, please comment on whether the expected effect size should be adjusted if an enriched dataset is used. If

so, how and why?

So again, to come back to the first -- can we come back to the first paragraph of this question? So if we're going to use an enriched dataset, how should we handle this and the different substrata of mammographic imaging?

DR. ROSENBERG: I'll take a shot.

CHAIRMAN GLASSMAN: Thanks.

DR. ROSENBERG: I think it's necessary simply because of the low prevalence that we use an enriched dataset. I think we run the risk when we start manipulating it to too great a degree that it doesn't become clinically representative data. So my suggestion would be that the enrichment be for cancers of the types that are routinely identified in clinical practice, and less manipulation of the size and types of findings that are present if that makes sense.

CHAIRMAN GLASSMAN: Dr. Tourassi, you had a comment and then Dr. Berry.

DR. TOURASSI: In principle, I agree with Dr. Rosenberg. Certainly, we need an enriched dataset. I do not believe in so much micromanagement of, you know, the size or the type of classifications. Definitely not palpable masses, that is a different group.

But in terms of how this dataset will be created, I think, it depends on if we're looking at the standalone performance or the reader-based performance because for the standalone performances we talked about the significance of having the different substrata. We certainly need to collect the sufficient number of cases for each one of the substrata.

But when it comes to the reader performance and this is actually what the system is going to do in the clinic, I believe consecutive cases which represent the overall screening population, this is what matters. And because, of course, we cannot collect enough cancer cases in a consecutive

population, I would say enriched dataset with consecutive cancer cases.

By having that consecutive collection of data points, we know that we represent the overall population, the overall prevalence of masses versus calcifications versus architectural distortions. So there should be different set of criteria, some more relaxed for the reader-based assessment of the devices.

CHAIRMAN GLASSMAN: Dr. Berry?

DR. BERRY: So I agree with Dr. Rosenberg about the representative cases, and I do think that enrichment is essential for all of the reasons that have been discussed. The dig on enrichment that we heard earlier, quite legitimately, is that the reader will come to know what the enrichment was and be biased accordingly.

And I take it that that's because the only thing that is being considered in enrichment designs is you randomize or you mix

up things, so that the -- there is a, you know, equal distribution throughout the entire study. We have talked about a database. You can imagine a database that you would enrich on the basis of cases.

What hasn't been considered and would be quite effective would be, not to randomize. You shuffle the cases, so that they are random, maybe, or worry about trying to get things consecutive, as Dr. Tourassi indicated. You shuffle the non-cases, but then you vary things, so that the proportion of cases is not always the same. That, sometimes you are in the background setting, where there is almost no cases.

And sometimes you have a greater proportion. And thinking forward where this database is going to be used for different companies, you would want to vary that from one to the next. So you have different functions that talk about that -- that relate the intensity of cases over time and that,