that function changes from one time to the next.

So you can't really predict. And I think that wrinkle on enrichment and the representative nature of the cases, I think, would make for quite a good study, but also a good setup for people to be doing multiple company products.

CHAIRMAN GLASSMAN: I apologize, I have to step out for five minutes. Dr. Mittal is going to take over for me and I'll be right back.

DR. ROSENBERG: Can I add one other point about the cases? We're focusing on cancer cases, that is important. But also, the cases that are not cancer, being representative, would also be important.

DR. BERRY: You mean benign?

DR. ROSENBERG: Yes.

DR. ZISKIN: I would imagine that the appropriate test series would be all the abnormal and the relative proportions that

exist in the population, just minus all the normals.

DR. BERRY: Not all of them.

DR. ZISKIN: Okay.

DR. MITTAL: Well, there was a comment from that side.

DR. SAHINER: Yes, I think the question is also asking about stress test, and I think it may be legitimate for some company to do a stress test. But in this situation, I think it should be clearly labeled that it is, you know, tested, or it has been designed for a certain kind of abnormality.

So if it's a stress test, I think it should be clearly labeled that it has been tested on a certain substrata of the dataset.

DR. D'ORSI: Yes, I think we're focusing on stress versus no-stress. It would be nice to mix it up. I think it should edge more on the stress and not eliminating the more obvious things or more blatant things that we see in clinical practice. To me, it's

sort of like a boxer training for a fight. They train with extra heavy gloves, so that when they go into really fight, they could move a lot faster.

And to me, it's kind of similar. I think that we should have a, perhaps, inordinate amount, not totally representing what's seen in the screening situation of stress tests, because when I or some of my colleagues look at this, they really want to know can I trust this thing when it says nothing.

Because if I could trust this thing with 99 percent probability, I'm home free. And then if I'm averaging one or a half a mark on the cases where I don't see anything, and I know it's really true, I'm going to be very relieved. That's not the case now. It's not the case, because we don't know where it falls down in this area, in these particular areas.

And the time to find out is with a semi-stress type of thing. If we just

duplicate the screening setting, I don't think we're going to stress it enough. We're just going to get data that is very similar to what we are doing now.

DR. MITTAL: We have a comment on this side.

DR. KIM: Yes, I would definitely agree with you in terms of standalone performance testing, that we should do kind of an enriched dataset in terms of stress. But I guess I would caution if we went to the reader performance test, because taking it from my experience in terms of colon screening, one of the things we have noticed with looking at different prevalence in terms of significant cancers and advanced neoplasia is that when the prevalence is high and you enrich the dataset, the reader can do very well from -- with what is not probably the optimal way of detecting cancers.

But if you use that same algorithm in a low-prevalence setting, the performance

markedly decreases. So, in a reader performance where you have enriched and stressed because of the abnormalities that are there so frequent, the reader may be able to do better and it may be more difficult to see the effect of CAD in that instance.

CHAIRMAN GLASSMAN: Dr. Berry?

DR. BERRY: Surprise! I want to comment on something that Dr. Sahiner said. He said that if we do stress tests, that we label them as such. And Dr. D'Orsi has suggested that we enhance the stress.

There is a comment here or question, "In addition, please, comment on whether the expected effect size should be adjusted if an enriched dataset is used. And if so, how and why?"

It should be adjusted if we -- well, first of all, I don't think that we should tell anybody what the test is. So I disagree with Dr. Sahiner. I also disagree with Dr. D'Orsi. I think that we should have

a representative set of cases enhanced or enriched, but a representative, otherwise, we have to answer “no” here.

If we have a representative set of cases that you know have some stress cases in them, but no more than in the general population, we're -- we do have an unbiased estimate of sensitivity and specificity, despite the enrichment.

And so we wouldn't have to make these adjustments. If we have the stress in there, then it stresses statisticians, because we've got to figure out when or what the adjustment is.

CHAIRMAN GLASSMAN: Yes?

DR. SAHINER: I want to clarify. I didn't say that we should tell the -- when we are doing the testing what the stress has been in the labeling. Let's say that you do your pivotal test with cases that were missed by radiologists, then I think in the labeling, it should say that this was tested on a dataset

where the cancers were missed by the radiologist because that's the stress that -- that's the dataset that you collected.

But you don't tell the radiologist when they are -- when you are doing the observer study that they were missed cases. So it's just the labeling that I'm talking about. And I had one more point that I forgot.

CHAIRMAN GLASSMAN: Well, we'll come back to you when you think of it. Yes, Dr. Dodd?

DR. DODD: I have two points that I want to make. One, I wanted to come back to this idea that Dr. Berry suggested about not necessarily doing randomized sequencing of the scans. I had a similar idea. Having some control of the sequence of the -- sequencing of the scans might also allow you to test whether there is some effect of the knowledge prevalence on actually whether they are likely to call.

Say if you have a sequence of positive exams followed by a negative, is the reader more likely to call that one a positive? So if you control the sequencing in some creative way, you might be able to test for that.

And then secondly, I just wanted to throw on the table how much enrichment are we talking about? Are we talking about going from a prevalence of .4 percent to 50 percent? Are we talking about -- should there be some limit on the amount of enrichment that is allowed?

CHAIRMAN GLASSMAN: Dr. Berry?

DR. BERRY: So we should add cases so that we have enough power to assess the sensitivity. And it's not the proportion of normals in cases, but the numbers of cases and the numbers of normal.

DR. DODD: Well, can I follow-up on that?

CHAIRMAN GLASSMAN: Sure, Dr. Dodd.

DR. DODD: Yes.

CHAIRMAN GLASSMAN: Then I want to take over.

DR. DODD: I'm not sure that that's the key. I think it is the ratio of the normals and the cancers in the sense that the radiologist begins to change their calls based on -- so I wasn't -- I mean, obviously, you need to power it, so you have enough cancers to detect a certain change in sensitivity, but how many normals do you throw in is the question I'm asking.

CHAIRMAN GLASSMAN: Let me go through the -- this question piece by piece, I think, to make sure that we cover all of the things. The first subpoint was breast density, 40 to 50 percent, dense breasts. Is that a good thing? It sort of mirrors what is clinically available. And since we know that cancers tend to be missed in dense breasts, is that something that we would like to suggest is included in the dataset? Any comments

about apple pie, motherhood, dense breasts? Dr. D'Orsi?

DR. D'ORSI: Yes, it depends on how you define dense breasts. If you define it the way ACRIN did, then I agree. ACRIN said the categories of almost entirely fatty and scattered would be non-dense. And the categories of heterogeneously dense and extremely dense would be dense.

And if you look at that breakout, it is about 50/50 and would hit exactly what your screening population is. So I think that should be done.

CHAIRMAN GLASSMAN: Anyone -- oh, Dr. Berry?

DR. BERRY: Yes, so just to comment again on the expected effect size and whether it is unbiased. If you enrich for cases, you don't have to make this adjustment. If you enrich for breast density, then it's not impossible, but you do have to make an adjustment in the effect size, based on the

enrichment for breast density.

CHAIRMAN GLASSMAN: Okay. But there is -- is anyone opposed to a 50 percent dense breast target as a good thing for evaluation of breast CAD? Okay. Let's go on.

Proportion and types of masses and microcalcifications, approximately evenly distributed with sufficient number of additional patients with architectural distortion alone.

So, a sort of even mixture of the most common causes of imaging-detected breast cancer. Is that a good idea, bad idea, needs to be modified? I'm sorry, Dr. D'Orsi. I keep looking to my left. I apologize to you people on the right.

DR. D'ORSI: I suppose you are a lefty.

CHAIRMAN GLASSMAN: No.

DR. D'ORSI: No? I think this is a very interesting statement. It gets back to something that Dr. Berry said. I think we

should add additional patients with architectural distortion and thinking about Dr. Berry's answer, I kind of agree with the calcifications and the amorphous.

But architectural distortion, if you miss it, is lethal. It only means two things. It means cancer or injury, period. So to ensure that that is properly labeled and to test at a relatively stress level for this, I think is extremely important.

DR. TOURASSI: I have one question.

CHAIRMAN GLASSMAN: Dr. Tourassi?

DR. TOURASSI: Are we talking -- are we clarifying standalone performance or reader-based performance? So all of these recommendations go under what?

CHAIRMAN GLASSMAN: What would you like them to go under?

DR. TOURASSI: No, because I echo what Dr. Kim said before. We should have different standards for the standalone performance, where we have the luxury of

getting cases, and different standards for the reader performance because with any kind of enrichment and any kind of stressing, all of us know that when we bring the observers in to do the studies, they are far more vigilant because we expect that there is more there.

So we don't truly get what happens in the real world where we anticipate 4 cancers per 1,000 screens.

CHAIRMAN GLASSMAN: Well, let me go then to the end, and that is, should w, therefore, expect a higher performance with an enriched dataset? And if so, how much higher?

DR. BERRY: Sir?

CHAIRMAN GLASSMAN: Yes, please.

DR. BERRY: You are going to the end? What do you mean the end?

CHAIRMAN GLASSMAN: Later on in the question, there was something.

DR. BERRY: In addition, please, comment on?

CHAIRMAN GLASSMAN: Yes.

DR. BERRY: So this is what I have said a couple of times with respect to specific cases. But all five of those Roman numerals are problematic in terms of introducing biases and have to be addressed. And they can be addressed, but I point out that they are counter to Dr. Rosenberg's first comment that we should have a representative set of cases.

So there is one approach which would be to take cases, enrich according to the cases, but you get equal proportions of masses and micros, et cetera. And if you're going to enrich on the basis of masses and micros and these other things, breast density, then you screw up the biasness, so that you have to make adjustments to get back to what the effect size is.

It's not impossible, but it changes the -- my notion of a simple study, which you know, I signed on to Dr. Rosenberg's statement that we should be representative. But you

know, I'm open about these other things. It's just it makes life difficult.

CHAIRMAN GLASSMAN: Dr. Rosenberg?

DR. ROSENBERG: Yes, it seems we are talking about reader performance which is a difficult testing procedure to do. And so, I think the concept of simplifying this part of it, where if you are looking at standalone, you could enrich easily with subsets because it's really just collecting the cases and having the computer do its algorithm.

But this involves a radiologist. This is actually a difficult thing. And under the concept of making it reasonable for the vendors to prove efficacy, it seems we do want to make it simpler. So not have all the subcategories necessarily enriched.

DR. LIN: I think it's going to be very difficult for the statisticians to adjust adequately when we introduce all this stress testing, especially when we are measuring reader performance, you know, outcomes. So I

think it's problematic with this. So I would probably be more in favor of case enrichment, but still in the usual proportions that we see in clinical practice. I think that's -- the stress testing really just introduces unnecessary complexity to the whole process.

CHAIRMAN GLASSMAN: Let me ask a questio, because I am now totally confused. I'm -- we have a disease with an extremely low prevalence. So if we're not going to have, you know, 40,000 patients in a reader study, we need to do something to enrich the population, the test set for the reader study.

I agree, for a standalone study, we can do as many as we need. We can enrich it. We can do whatever we want. But for the reader study, for the readers not to simply see all normal mammograms, we have to enrich the dataset somehow. Enriching it in proportion to the disease absent the normal patients makes to me, some sense. And then throw enough normals in to make it reasonable

to adjust the statistics in some way that you statisticians do.

Is that not where we want to get or, I'm not trying to drive the answer here, but I'm totally confused.

DR. BERRY: So that is the simple version that Dr. Rosenberg suggested, and then we started talking about enrichment on the basis of cases. We then launched into a discussion of the important roles of breast density and micros versus masses. That the Roman numerals i through v, these are different types of enrichment.

And so, if we enrich on the basis of cases, but also enrich on the basis of these characteristics, then we have problems. They are not insurmountable problems, but they create lots of difficulty.

So the question is should we be enriching on the basis of characteristics other than cancer? And at least some of us think that it's better to -- I think you do,

as well, Dr. Glassman, that we enrich on the basis of cases, and let these other things fall the way they do normally.

CHAIRMAN GLASSMAN: Well, again, I don't want to make this a two-way conversation, but I'm going to one more time. My comment: and that is what I'm envisioning is that the enrichment is i through v. but in a ratio that respects the normal incidence of those things in a non-normal population.

If you took a thousand cancers, you would have so many clustered calcifications, so many masses, so many architectural distortions; that's the ratio that I'm envisioning. Is that a statistical nightmare for the statisticians?

DR. BERRY: No, you are doing it in a simple way, and these characteristics are in the back seat. What is driving it is the cancer. You are putting it in the enrichment because it is cancer, and because it's cancer it happens to be, you know, more likely to be

breast dense -- dense breast, et cetera.

So these come along for the ride in your approach; whereas, you could think of enriching more on the basis of these, both in the normals and in the cases which is problematic.

DR. ABBEY: Could I just ask another question then? So if we are enriching, we heard this morning that about 10 percent of all cancers are not visible mammographically at all, how do you enrich with those? You don't even know if the cancer was there because you detected it a year or two later. And you don't know if it was present at that time.

So how do you enrich occult cancer cases? I just see a structure -- I just see the idea of an unbiased estimate of sensitivity by any sort of enrichment as kind of difficult to attain.

DR. ROSENBERG: Actually, it's a good question. I would suggest enrichment

just be a detected cancer is -- detectable cancers, but that becomes a difficult question.

DR. ABBEY: So that's a verification bias then, that's in the thing, because they were detected cancers.

CHAIRMAN GLASSMAN: Carl? I guess, Dr. Ziskin.

DR. ZISKIN: There was a comment made about a concern that if a reader knows that there has been a stress thing, it would change the behavior. And of course it will, however, my understanding is that the ROC analysis will take that into account of this. That that's the purpose, I feel of the ROC.

DR. TOURASSI: Not necessarily, because if they have become more vigilant and they detect more things because they go back over and over again, all the benefit we get from CAD in terms of the perceptual letter is going to be lost under this scenario, so the ROC difference.

DR. ZISKIN: Maybe at the time.

CHAIRMAN GLASSMAN: Let me ask another question then. There was another part, I think, to this question that we didn't project.

(b) If you believe that enrichment is appropriate, please provide your reasons. That we have done. And whether there would be an alternative method. It should be inappropriate, I think, there.

If you believe that enrichment is inappropriate, please provide your reasons and whether there would be an alternative method of assessing these devices in light of the low prevalence. Is there something we haven't thought of? No? Okay.

I would like to, unless somebody has a burning comment -- Dr. D'Orsi has a question.

DR. D'ORSI: Just one thing, and this just hit me. I agree now with what Dr. Berry is saying. I can understand the problem

of enriching with a performer, but you don't disagree with doing that on a standalone, the stress? Okay.

The other thing; this is interesting. When you look at a mammogram, you are thinking of how much this would indicate that it is malignant before you call it back. But there are many things. In other words, the call back, you may not necessarily be thinking 100 percent malignancy. You may be thinking this has a chance.

So there are a lot of benign things, and that's a problem with mammo, that simulate the malignant findings. How do we capture that? Is it possible to capture that, or is that just too complex to start mimicking findings that are pathology-benign, or is that just a nightmare?

CHAIRMAN GLASSMAN: Go ahead, Dr. Berry, and then I'll summarize.

DR. BERRY: Yes, I would say it's a nightmare. It's the -- you know, we hear

these things all the time that HRT is responsible for cancers that are worse prognosi, because it's really difficult even in the context with lots of data.

I do want to point out this comment about the occult cancers. And Dr. Rosenberg said that we should do detected only, but that leads to bias. So with respect to this additional comment, if it's possible, the cases should be cancers. Cancers that are identified at some future time, whether or not the cancers were visible at the time is -- maybe there is a CAD that would find the thing that we didn't think was visible.

It should be cancers. If it's not cancers, I would just point out -- I mean, if its detected cancers, I would point out that that's a bias and has to be adjusted in some fashion.

CHAIRMAN GLASSMAN: Let me try to summarize where we are, and I hope my own biases don't creep in too much. If they do,

please let me know.

That for both standalone and reader testing, the Committee recommends an enriched dataset. For standalone testing, the enrichment can be stressed and can be a large set because we are not worried about reader bias.

For the reader sets, however, the stress should -- I mean, not the stress, but the enrichment should mirror the normal distribution of cancers in the imaging population with, if possible, prior "negative mammograms from patients who developed cancer that was visible a year later, if that's available."

All of the categories we agreed with the different kinds of calcifications, the densities, the small masses, and that we did not believe that there was an alternative to enrichment.

Ms. Brogdon, does that answer the Agency's need?

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

CHAIRMAN GLASSMAN: Great. Thank you. Okay. The fifth question, we have 5, 6, and 7, and then we have a coffee break to give you something to look forward to.

M5. Mammograms obtained on Full Field Digital Mammography devices have characteristics that are strongly dependent on engineering design and device hardware and software. If a mammography CAD has been approved to operate with screen film or a specific full field device or devices, what data should be used to assess its performance with the different Full Field Digital Mammography device?

Would one or the other suffice? Are there other types of studies that should be provided instead of or additionally to?

As we're going to full field digital more and more in the United States, I think this becomes a very important question.

Just let me, for those of you who don't know as much about full field digital, the way that it works is that the image is taken and stored electronically. It is then manipulated so that the end result looks like a more beautiful than beautiful film mammogram.

But there are steps in the middle where there is data manipulation that are machine-specific, company-specific, and probably not necessarily generally available as trade secrets.

So one of my concerns is that, if we say that -- if the companies look at the final imaging output and that's where they put CAD, then they are, in effect, CADing an electronic film mammogram. If, however, CAD comes in at a much earlier stage, then it really may be machine-specific in its workings. That the manufacturer of CAD may have to deal with the manufacturer of that device.

And so there are several levels to

this question that I think we need to talk about. With that, I will open it up for comments. Dr. D'Orsi?

DR. D'ORSI: The only time I don't raise my hand you call on me. It's a great question. I was trying to think of it, but -- think about it. The -- I don't know enough of what goes into process the raw images to know if doing a CAD on the raw versus the processed images will make a big difference between manufacturers. I just don't have a feeling for that.

I would imagine that, just guessing, most of the manufacturers have a pretty similar processing algorithm. So intuitively, you would think it wouldn't make a big difference, but I don't know how to answer that.

CHAIRMAN GLASSMAN: Dr. Berry and then Dr. Garra.

DR. BERRY: So it is conceivable that the digital mammography has greater

accuracy and that it works differently with readers when you are using CAD. So it's conceivable that the CAD efficacy is now less or even gone in the context of digital films.

The -- I would suggest that you would have to do a study that shows that in the context of this new full field or full field -- if you are approved for mammography, for film and for full field, you would have to do a new study. You ought to be able to borrow, to some extent, your -- from the historical study, from the historical data and not have to do as big a trial or as big a study as you had before.

But I think the onus is on you to show that you are as good in the context of full field as you were in the context of film.

CHAIRMAN GLASSMAN: Dr. Garra next and then --

DR. GARRA: I do think that there are going to be differences between the systems and that a specific CAD system may be

tailored for the specific output of a full field digital system or a CR system or whatever. So I think there will have to be testing, but it may be possible to get at those results by doing a standalone test, rather than an observer trial.

CHAIRMAN GLASSMAN: Dr. Spindell?

DR. SPINDELL: That was exactly my question was whether Dr. Berry would feel a standalone trial in that situation would be sufficient?

DR. BERRY: No.

DR. SPINDELL: And what trials would you suggest?

DR. BERRY: Well, the reader trial, the adjunct trial, because it really -- what matters is, you know, what the reader can see in the full field as opposed to the film. And that could be very different. That is, the CAD may be performing exactly as it had in the standalone, but now, you don't need it because the reader can see in the digital things that

the reader couldn't see otherwise.

DR. SAHINER: Maybe if more knowledgeable people about the ACRIN Study might correct me, but ACRIN Study on a large population of cases comparing screen film to digital, for a large majority of them, there was not a statistically significant difference between screen film and FFDM. It was only for younger patients, pre-menopausal women with dense breasts.

So then the question is, for example, would it be sufficient to look at only that subpopulation where we see a difference between FFDM and screen film?

And the other thing is the other difference between screen film and FFDM is, of course, that the FFDM has been processed. So it's different from the issue of whether the lesions are better visible on FFDM or not. But different companies may have different pre-processing algorithms. And if you are looking at the end image, then depending on

how you do your processing, the results that you get might differ, based on, you know, which company's pre-processing you are working on.

So I think for the second question, you know, different companies pre-processing algorithms and how your -- how a manufacturer's CAD works on that, I think for that one, a standalone study might be enough, because it's not a, you know, perception or it's not a user issue. It's how your CAD works with that company's pre-processing algorithms.

CHAIRMAN GLASSMAN: Any other? Dr. D'Orsi again.

DR. D'ORSI: I should never do this. No. What was I going to say? On the processed images, I think I agree with Dr. Berry that a performance reader assessment is even more critical than it is with film screen, because of the processing.

What I have anecdotally noticed on

these images from the manufacturers is they look beautiful. They are gorgeous. You can hang them in a museum, but I don't want to hang them in a museum, I want to find cancer. And what happens is that much of the anatomic detail, the anatomic noise is also enhanced.

So now, your eye is bombarded by anatomic structures that are enhanced almost as much as what you are looking for and now your head is trying to say oh, my God, I have a million criss-crossing lines and densities. My God, they are all standing out. And that's a significant effect with the processing.

And I think that the performance testing, the reader testing are going to be very critical on that to assess the level of processing.

CHAIRMAN GLASSMAN: Dr. Garra?

DR. GARRA: I just wanted to say that I agree that there could be differences in human observers with full -- various types

of processing in full field digital, but that's an issue for the FDA to address in performance of Full Field Digital Mammographic systems, not CAD systems.

And what we want to establish here is whether the CAD system is the same, and if there is variations in full field digital, that's a different question that the FDA may need to address.

CHAIRMAN GLASSMAN: Dr. Watt?

DR. WATT: Only processed images of the FFDM, the final appearances depend -- are vendor-dependent. And so I think it's essential that there is another reader study, because the standalone would convert, yes, whether the CAD does or does not operate adequately, or to the same standards it did with film screen.

But the differences on the final processed image are such that, indeed as Carl says, you have such anatomic detail and the detail that you have from the cancers as well,

I think that you need again to see how the CAD and each of them are operating, and that they are operating equally well vendor to vendor.

CHAIRMAN GLASSMAN: A comment? Yes, Dr. Bourland?

DR. BOURLAND: So if they -- the two systems are a single system and coupled such that the data from one funnel directly, than clearly that's a system that needs to be tested at both ground-level standalone, as well as reader. If they are independent, I agree with recent statements that they -- that is no different than what we were already talking about, about a standalone CAD system that now has an image input from somewhere and that needs to be tested at both standalone and reader levels.

CHAIRMAN GLASSMAN: Are there any other types of studies that should be provided with full field digital either in addition to or instead of standalone and reader testing?

DR. SAHINER: Can I go back one

step and make a comment? Now, studies like the ACRIN Study, they are trying to look at the observer performance alone, radiologist's performance alone, FFDM versus screen film. So if those two are about the same, and if you have a CAD system that works about the same on an FFDM versus screen film, then I think it is, you know, splitting hairs a little bit too much to then say okay. But when they are combined, although they are in the same, when they are alone, when they are combined, there might be more errors, so let's do another observer study.

And I would like to remind that there are many different FFDM systems by different companies with different pre-processing. So that would make a lot of new observer studies for each of them. It would be, I think, very burdensome.

CHAIRMAN GLASSMAN: Does anyone want to change their mind with a comment? It seems that the sense of the committee with one

very cogent objection is that both reader and standalone testing would be necessary. Is there a --

DR. TOURASSI: I would like also to echo Bourland's comments. I do agree with him on that.

CHAIRMAN GLASSMAN: Okay.

DR. ROSENBERG: Yes, I think with the multiple CAD or new digital systems out there, and digital systems also will evolve. In other words, what we are with now is a first generation digital so if we suggest the reader performance testing for the first generation of digital, where are we when the detectors change? And at what level do we require a new observer performance for the digital machine, not just the CAD?

So I think we are in -- that could be a very difficult burden.

CHAIRMAN GLASSMAN: Yes, we are here, though, dealing with CAD, rather than with a new digital machine, except as an input

device into the CAD.

DR. CARRINO: I would also augment that the CAD is accepting a digital image and you, as a -- you know, either through MQSA or whatever routine clinical practice you have, you have defined what is a diagnostic clinical image, and that's the input to the CAD. So whether it is from one of these full field digital machines or something else, I'm not sure that that matters.

I think it is too burdensome to do it for every single type of vendor who makes those devices. So it should just be looked at as a digital image coming in and the CAD is testing it. Now, it would be -- it probably would be prudent and behoove the vendors for CAD to look at a range of these different devices that exist out there.

But I think to mandate it would be too burdensome. If there is a CAD product that is inherently involved in the image processing then, of course, if that's linked

to a specific product, then that's only good for that one system that it is tied into.

CHAIRMAN GLASSMAN: Dr. Berry?

DR. BERRY: So I think you have a split decision, Dr. Glassman. And just to be clear, my understanding of what we're about here is not is CAD the same from one to another? The uncertainty in this is the reader, and the reader can have a different performance in one setting than another. Just because the performance of full field is the same in the ACRIN Study or, you know, within statistical limits, doesn't mean that adding CAD is going to be the same in both cases.

So it's not is CAD the same from one system to the other but what does CAD add from one to the other?

CHAIRMAN GLASSMAN: Okay.

DR. GARRA: I have to reiterate that if the CAD is performing the same from one system to another and the reader is varying with the full field digital from one

to the other, that's a full field digital issue, not a CAD issue. To place that burden on the CAD developer or company, I think is unfair. You can't do that. It has to be addressed to the -- by the FDA as a variation in full field digital.

CHAIRMAN GLASSMAN: I'm still confused as to this -- well, I still hear a split decision. Dr. D'Orsi?

DR. D'ORSI: I now, clearly now, understand what Dr. Garra is saying. There might be variation. There will be variation in interpretation, but to dump it on CAD is unfair. As a matter of fact, CAD may actually help overcome that processing in a way. So I clearly see what you are saying.

CHAIRMAN GLASSMAN: Dr. Dodd and then I'm going to pose a question to the group.

DR. DODD: Okay. I just want to make a slightly different point. And that is, we're all assuming that CAD is going to

perform the same with the full field digital as with the screen film. And based on what I have heard, the way the features look and appear, it's not clear that they will behave the same.

So I think a question is if you have say a similar sensitivity but a decrease in specificity, in my opinion, that should definitely trigger a reader study. Any time you get an increase in false positives, you need to understand how the reader is going to interact with that.

CHAIRMAN GLASSMAN: I would love to be able to summarize what we just said.

DR. GARRA: I just wanted to comment on that.

CHAIRMAN GLASSMAN: Go ahead, Brian, and then I'm going to have to go on.

DR. GARRA: The assumption was that the CAD system performs the same on the standalone study. If it doesn't, then all bets are off. Then you are going to probably

have to do both.

CHAIRMAN GLASSMAN: Ms. Brogdon, the sense of the committee is, I think, that we don't know enough about the vagaries of image processing and the link between CAD and full field digital. That if the link is such that it looks like CAD is late in the process, standalone testing might be enough. I may take hits for this.

However if it's early in the digital process, where we can't be comfortable that it runs like film, that reader studies may be needed. Before you answer whether that is sufficient, is the Committee willing to go with that statement or do I need to modify it?

DR. BERRY: I think you just defined a unique position between the two extremes. And you said the Committee doesn't know. I think the Committee knows, it's just they know two different things.

DR. SAHINER: May I make a comment?

CHAIRMAN GLASSMAN: Yes.

DR. SAHINER: I think actually earlier in the process before the manufacturers' pre-processing algorithms are applied, the images between different FFDM manufacturers may be more similar. It's the processing that they do that we don't know about, and that's the variation.

CHAIRMAN GLASSMAN: You may be right, although I would like to suggest that your first part of your statement, at least to me, is an assumption, rather than a fact. I don't know that that's the case. Again, let me ask the committee whether you are willing to go with splitting Solomon's baby or not here. And then we will pose it to the Agency as to whether they are willing to accept that as an answer. Ms. Brogdon?

MS. BROGDON: I think you probably gave us more than we had hoped for. It was a heroic effort. Thank you.

CHAIRMAN GLASSMAN: We're going to move on then to M6. We have M6 and M7.

M6. Mammograms obtained on full field -- no, that's 5. Please, I don't want to read that one again, Sunder.

Here we go. FDA does not specify indications for use but reviews indications for use that are requested by companies. What are the Panel's views regarding second reader versus concurrent reading using a CAD device? Specifically,

(a) How are mammography CAD devices used clinically?

(b) Are second reader and concurrent reading modes both clinically relevant options for use in practice? If not, which paradigms are appropriate for mammography CAD devices?

(c) Do you believe users understand that if a device is labeled as a second reader, that is, the physician should always read the radiological image completely before turning to CAD?

Okay. That's the whole question. I would like to start off this one if I can.

I think, clinically, CADs are used in mammography, both in the concurrent and sequential mode. That's based on personal observation of mammographers working in the field.

I think we have been told, and I think it's true, that film reading is faster in the concurrent mode than in the sequential mode. And that has become useful to a lot of people with their workloads. So I think the way it is used is both.

Are they both clinically relevant? I believe that they are. The danger, of course, with the concurrent is that the concurrent reading becomes the first reading, and I -- without a complete second reading. And I think if a radiologist has the discipline to use it as a concurrent method, you don't lose any accuracy and you may gain time.

If you don't have that discipline, then with breasts you have to do it as the CAD

second.

And third, I'm really going to show my ignorance here, but it was my understanding that the labeling did not preclude practicing physicians from using the device as a concurrent reader.

Any other comments, please, from the Panel? Dr. D'Orsi?

DR. D'ORSI: Let me go through these three also. I think you are right. I think they are both used concurrently and as a second reader, but I think there is enough data to really, really raise a lot of red flags about using it concurrently for many reasons, several of which you mentioned. The bias of having something labeled as you are reading.

Studies have shown that true lesions, not marked, are totally ignored. And we are just introducing another layer of uncertainty by reading concordantly. So I think that's really something that is very

frightening to me, and I don't believe that everybody understands that it should be used as a second reader.

As far as the time is concerned, if they are using it correctly, in my opinion, correctly being as a second reader, you are going to take more time to read it, period. And that's one of the things I was suggesting to time these things.

If somebody reads faster with it, I think you have to be a little bit worried about what they are doing. So anyway, those are my opinions.

CHAIRMAN GLASSMAN: Dr. Watt?

DR. WATT: I'm in agreement with Dr. D'Orsi. Most of the time the clinicians are not aware or do not read -- they should be using it as a second reader. And my suspicion is, anecdotally, that there are clinicians who are using it simply as a primary read, are flipping on the CAD and zipping through cases.

There is no way that you can

decrease your time of reading with CAD at all if you are using it correctly. You have to be increasing your time.

CHAIRMAN GLASSMAN: In terms of these specific answers, I take it you think that the labeled use is the correct one?

DR. WATT: Yes.

CHAIRMAN GLASSMAN: Okay. Dr. Leitch?

DR. LEITCH: I think this is going to be the tension we're going to deal with in talking about all these CAD devices is this concept of the time element in a screen that could be saved by the computer and then focusing on the lesions at hand.

However, I think in breast, it is a circumstance where the difference among the individual lesions is probably sufficiently complex that this is more -- and you don't -- and you have a smaller field to look at. You don't have the whole colon or something like that, that you are probably better served to

use it as a second reader.

And, of course, if what you are saying to patients, and this is, to me, again honesty in practice, if you are saying to patients that, by using CAD you are giving them a second read, then that's how you should practice it. You know, if you're saying you're going to give them a better read instead of another person, you're having a computer give them a second read, then that's how -- you should practice it in that fashion.

On the other hand, if what you are saying is, you know, well, I got these kabillion people I've got to screen, and I'm going to have to start practicing some efficiencies, then that -- this concurrent might be something you would do, or somebody that's a really great reader could work it well.

But I think for most cases it should be the second reader practice, that's how it should be done.

CHAIRMAN GLASSMAN: Dr. Rosenberg, any comments?

DR. ROSENBERG: Yes, my experience would be it's -- there are people using it more in selective cases concurrent, I've been told for like fatty replaced breasts. You can pick out the calcs quickly with the CAD.

My personal experience is the way it is labeled, and I think that's the appropriate way to use the device.

CHAIRMAN GLASSMAN: Any other comments from any members of the panel? Okay. Ms. Brogdon? Oh.

DR. STEIER: No, I was just going to say there is a kind of parallel universe. I read EKGs and pulmonary function tests. And with EKGs you get the automated interpretation and highlighted areas of, perhaps disease, and the same thing on pulmonary function tests where the tests come back abnormal from the computer with the areas that are abnormal highlighted.

And over time, what I have noticed is people become very used to using the automated responses and become less and less diligent in reading things themselves. And in radiology, I don't know if that's true for you guys who read the mammograms and such, but I know in other areas of medicine it happens.

CHAIRMAN GLASSMAN: Okay.

DR. STEIER: And I don't know what the labeling says for either product.

CHAIRMAN GLASSMAN: Well, Ms. Brogdon?

MS. BROGDON: That anticipates my question to the Panel. What can FDA do to enhance users' reading and comprehension of labeling?

DR. ROSENBERG: Reading or comprehending?

MS. BROGDON: Both.

DR. ROSENBERG: Right, or utilizing.

CHAIRMAN GLASSMAN: Let me

summarize, and maybe I'll expand a little bit on that and see if the panel will go along. How are they used clinically? The answer is both concurrently and sequentially. Are they both clinically relevant options? I think the answer is that secondary reading is certainly preferable based on the data and the labeling that the FDA has.

However, under certain circumstances, concurrent may be acceptable but certainly not the best practice. In terms of how this can be made more -- you know, better to the physicians, I think the training from the companies and we haven't mentioned this, but if anybody disagrees, I'll give you a chance to have a second go at it, but certainly training from the companies as to the best way.

And the company is making the users aware of the data that shows that it is the best way would probably be most effective. Does anybody disagree with that or have an

alternative? Is that an adequate answer?

MS. BROGDON: I think the question is -- first of all, you need to understand FDA does not regulate off-label use of devices. And part of the question for you is whether the devices are being used in line with the labeling and the indications for use at all.

If you feel these devices are being used frequently off-label, that is for concurrent reads, do you believe that the labeling should eventually conform to FDA's approvals or clearances can conform to the real use of the devices? And if so, what data would be needed to get approval or clearance for the concurrent reads?

CHAIRMAN GLASSMAN: Dr. D'Orsi?

DR. D'ORSI: I am going to go the other way. You have more evidence to suggest not to read concurrently than you -- there is none that I know of to read concurrently. So maybe you should be thinking this should not be used and, you know, some stronger language.

There is stuff -- data, as I said, that indicates it should not be used concurrently, only because the satisfaction research and missing lesions that aren't circled.

MS. BROGDON: I think we are talking here about future studies. If a company wanted to obtain approval or clearance for indication as a concurrent reader, how should those studies be done?

DR. TOURASSI: Another reader study, based on that paradigm.

CHAIRMAN GLASSMAN: Right.

DR. GARRA: And that is the wave of the future. It's going to have to go in that direction because of time pressures in everybody's lives. So the companies should be warned that you are going to have to move in that direction, and they are going to have to move and improve their algorithms to go along with it.

CHAIRMAN GLASSMAN: So a reader study comparing secondary reading to

concurrent reading that would have sufficient statistical power to prove or disprove the premise would be the answer, I think, to your question?

DR. TOURASSI: Why necessarily concurrent reading versus second reading? The whole idea is the new paradigm, whatever it is, versus the standard way of no CAD. Because there is no concurrent reading and down the line there may be other paradigms.

CHAIRMAN GLASSMAN: I see nodding of heads all around the room. So I correct my statement. Concurrent versus standard. No more nodding, that's good. Is that a sufficient answer to M6 for you?

MS. BROGDON: Yes, thank you.

CHAIRMAN GLASSMAN: All right. M7, the last mammography question. FDA has provided you with a bibliography, that is us, of published literature for mammography CAD. Please, discuss whether these publications provided us with any additional information as

to how such devices should be evaluated in the future.

That was the CD that we all got. Any comments after reviewing the papers? Mine on reviewing was that I thought that it added a lot of postmarket information, but wouldn't -- didn't give me any insights into more effective premarket look at anything.

Anybody else have a comment?

DR. GARRA: Agreed.

CHAIRMAN GLASSMAN: That was a quick one. The answer was no. Are there any other -- would you like -- Ms. Brogdon, the answer is the data did not give us any premarket hints as to more effective ways to look at the equipment.

MS. BROGDON: Let me just ask the staff if there is a follow-up question. We're fine. Thank you.

DR. GARRA: Can I just make a comment on that?

CHAIRMAN GLASSMAN: Of course,

Brian.

DR. GARRA: Because that went too quickly, right, so we have to create some controversy. My reading of a lot of the postmarket studies did show the problems you can run into by trying to get statistically significant changes when clearly the changes-- the marketplace is showing the changes are people find it useful, but a lot of the studies aren't showing much in the way of statistically significant changes, particularly an area under the ROC curve.

And so it's a cautionary note that you could set the bar too high by using some of the criteria in those studies as a premarket device to require them to meet before they could get approved.

CHAIRMAN GLASSMAN: Okay. Any other comments? If not --

DR. BOURLAND: A comment.

CHAIRMAN GLASSMAN: Oh, comment, I'm sorry.

DR. BOURLAND: So I think some of the issues might have been in one of the presentations relative to dynamic studies 4D, temporal changes, these types of things, other modalities than x-ray.

CHAIRMAN GLASSMAN: Yes. We are going to get to that tomorrow. That is, if we ever finish today. If no one has any additional comments, we will now conclude our discussion of mammography CAD devices. We will now take a 13 minute break and come back at 3:35, please. Thank you.

(Whereupon, the above-entitled matter went off the record at 3:23 p.m. and resumed at 3:37 p.m.)

CHAIRMAN GLASSMAN: We will now proceed with the FDA presentations highlighting current issues related to colon CADs. Our presenter will be Dr. Frank Samuelson, from the Office of Science and Engineering Laboratories.

DR. SAMUELSON: Thank you. I'm

going to talk to you today about colon CAD devices, and some of the issues related thereto. First, here's my outline. First, I'll discuss briefly about colorectal cancer, the diagnosis and care thereof, including, specifically, optical colonoscopy and computed tomographic colonography.

And then I'll move into CTC computer-aided detection and diagnosis, or colon CAD, as I'll refer to it. And I'll talk about the potential of some of the details of the implementation, performance studies, standalone and reader studies, and issues particular to CTC CAD studies.

So colorectal cancer accounts for about 52,000 deaths per year in the United States, so that's why we're here this afternoon specifically. Colon cancer is believed to arise primarily from adenomatous polyps. Most polyps, 90 percent of them or so, are -- tend to be hyperplastic, not adenomatous. And these are mostly benign.

Larger polyps are -- have a greater -- the larger of polyp, the greater the likelihood that it is -- might be an adenoma or a cancer. About 10 percent of adenomatous polyps larger than 1.5 centimeters may contain invasive cancer. About 8 percent of the screening population has a polyp larger than 1 centimeter.

Fortunately, these polyps are relatively easily removed, and that's why colorectal screening is recommended for people over 50 years of age.

Methods for diagnosing colon cancer and finding polyps include barium enemas, in which case barium acts as a coating of the colon, and as an x-ray contrast. There is also fecal blood testing, which is inexpensive, and very safe and easy, but tends not to be very specific nor sensitive to colon cancer.

There is also optical colonoscopy, which is a current screening standard. And in

an optical colonoscopy, there is a series of bowel cleansing, there is insufflation of the colon, and then an endoscope is used to view the inside of the colon and visually search for polyps.

Relatively newer, there is CT colonography, which involves some of the same steps as optical colonoscopy, such as bowel cleansing, and insufflation, but instead it implements two CT scans, one in the supine position, and one with the patient in the prone position.

Her,e I'm going to compare optical colonoscopy and CT colonography in a little more detail. The optical colonoscopy tends to be a bit more invasive than CTC, and thus requires patient sedation. But on the other hand, it requires no x-ray dose, unlike a CT colonography. And as I pointed out before, one uses a fiber-optic scope, and another uses a pair of CT scans.

Another very important difference

between the two is that optical colonoscopy provides the ability to perform a resection, or in this case, probably a polypectomy. Whereas in CT colonography, any positive finding, or a desire to do a resection, requires an optical colonoscopy.

All right. So in a CT colonography, the supine and prone scans are rendered on a digital work station, and they can be rendered as a 2D slice view, as you can see here on the right side in this image from a paper by Bogoni. There it is. It can also be viewed -- the data can also be viewed as a 3D surface or a fly through, which is often called virtual colonoscopy, and that's this view right down there.

And after doing -- after examining a CTC, a clinician will often give a report, which will include polyp sizes, locations, morphology, perhaps a summary diagnostic assessment for the patient. One proposed way of doing this reporting is the C-RADS system,

by Zalis, 2005.

This C-RADS system also includes a rating scale system, which has five different levels, zero through four. The zero level, in a method very similar to BI-RADS is -- essentially says, we need more information. It's an inadequate study.

Level 1 indicates normal findings, or benign lesions. And from there, the higher levels indicate greater likelihoods of malignancy where this malignancy is -- the malignancy is often -- the possible malignancy is often inferred from the size or number of polyps.

And so, for example, C2 says, for intermediate polyps, less than three in number. And where C3 may be polyps with -- may be a polyp larger than a centimeter, or more than three polyps of smaller size. And C4 maybe -- is essentially saying it's very likely to be malignant, and surgery may be necessary.

The performance of CT colonography has been studied in a number of papers. Kim and Pickhardt papers indicate that neoplasia detection rates are similar to optical for polyps greater than 10 millimeters, and Cotton has found -- but Cotton found a lower detection rate for CTC than optical.

What is very true is that, in almost all cases, detection rates are strongly dependent on polyp size. And here -- and the chart below shows some results from the National CT Colonography Trial. And you can see that behavior. The specificity remains relatively constant across all polyp sizes, but the sensitivity rises strongly with size.

Along with CT colonography, there are now, of course, CAD devices. And these CADs work -- function in ways similar to other CADs. They mark suspected polyp locations on a CT work station display, and then the -- it is left to the doctor to determine whether each mark is significant or not.

These marks can be made on either a 2D rendering, or 3D rendering of images. And here there is two examples, one from Halligan, and one from Summers, on the right side, in two and three-dimensional views, respectively.

The potential benefits of CT CAD. One, of course, is improved polyp detection. Two, the second one may be reduced reading times, and three might be -- well, it goes on hand in hand with No. 1 in that it may provide recommendations or guidance for optical colonoscopy.

And you will be -- and these refer -- these points refer to discussion question C1, which you can kind of see popping over here on the side again. And so we will be asking you, the Panel, to give us input on whether these are valid potential benefits, or if there are any where we happen to be missing.

I'm going to quickly step through the possible implementation of a CTC CAD in a

very general way, rather quickly. You will find that there are probably similarities with other types of CADs, as you heard in the general CAD section earlier this morning, and, but do note that CAD devices will vary significantly one from another.

So in general, the first step of a CT CAD is to segment the image. And the image -- so the voxels of a CT scan are grouped corresponding to organs. And in the image you see on the right side there from Pickhardt shows the cross sectional -- a 2D cross section of a CT scan. And as you can see, there is the white bone corresponding to the pelvis, and then there is also the large black areas, which are the insufflated colon.

And in the colon, you can see some remnants of oral contrast taken by the patient. And that's why they are bright white. The edges of these organs are detected, and they are used to create surfaces or meshes in the -- for example, the mesh at

the bottom, by Zhang, is actually a mesh created from a Phantom.

So then, once the surface of the colon is digitally created, the software goes and calculates the sphericity and curvature at every point on that mesh or surface. And in general, these algorithms tend to be fully three-dimensional.

Vertices with similar correlated curvatures are grouped together to create regions of interest. And multiple features for each of these regions of interest are calculated, and they're fed to pattern recognition algorithms, such as the ones listed there. And then the output of the CAD algorithm is used to decide which regions of interest should be marked for the user.

And in this example from the Taylor paper, on the right side, we see a CAD indicating a region of interest, or a polyp in this case, in both a two-dimensional view and a three-dimensional view.

So now, I'm going to speak about CT colon CAD performance studies. And so, I'm going to discuss two kinds of studies broadly. One is standalone performance studies, how well a CAD performs by itself, and the other one is reader performance studies, how well do readers perform when using these devices. And these will relate to discussion questions three and four.

And so in various -- these are very similar to the questions that you've had from the mammography section, as well.

So, for standalone performance studies, we're wondering how well a CAD device performs by itself. Most CAD devices will mark multiple locations on the colon that are not truly polyps, and thus, it is up to the doctor to dismiss most of these. And in general, we'll use -- and often times, in literature, FROC curves are used to demonstrate the performance of these devices.

And an example curve is shown up on

the chart here from Summers where he demonstrates multiple FROC curves of marks of a colon CAD's performance on different size polyps. So just to remind you, the vertical axis here is sensitivity, and the horizontal axis is false marks per patient.

Some -- these CT CAD devices generally have been -- what do I say, generally have had their sensitivities set down at around 90 percent. And these measurements are given for polyps greater than 10 millimeters.

Issues with CT CAD standalone performance studies are similar to other CAD devices that we are looking at. Specifically, I'm going to bring up the points of marking and scoring. For example, if a CAD marks a polyp in the supine scan but not in the prone scan, is that considered a true positive, or is it considered both a true positive and a false negative?

What overlap criteria should be

required for CAD marks to count a polyp as a true positive? Now, unlike mammography, the overlap criteria here will need to be fully three-dimensional. In which case, we have a volume which is a polyp, and then we have the -- and then we have a CAD mark which is supposed to be indicating that volume.

So now, I'm going to speak a little bit about retrospective reader studies. In other words, we want to know, in this particular -- in these studies, the idea is to understand, how is radiology performance affected by a CAD device? And these are related to questions C4 and C6.

Multiple reader -- these studies generally tend to be multiple reader, multiple case studies. And typically again, these case sets in these studies are enriched with polyps. And that's actually question C6, in which we will -- and we asked the Panel about the appropriateness of doing so.

Here is a number of retrospective

CTC CAD reader studies. They've got four listed here, along with the number of cases involved in each and the number of radiologists. All of these studies have demonstrated increased reader sensitivity, and they all demonstrate a decreased reader specificity.

The Petrick and Taylor papers found no significant change in the AUC of the readers.

Now I'm going to address a couple of issues specific to CTC CAD reader studies. A number of these readings and these markings can be done in either two-dimensions or three-dimensions. In other words, the primary reading, how the radiologist primary reads the study, can be done in one of two different ways.

And then, which opens the question to, are there -- are CAD devices more effective in one mode than the other? Additionally, there's the question of reader

modes. And when I say reader mode, I'm talking second reader versus concurrent reader. And this is another question we will repose to you as question C7.

And just to reiterate, the second reading mode, the clinician reads without the CAD, and then makes -- records their findings, and then they read with the CAD, and then re-record their findings. Whereas, in concurrent reader mode, the clinician just switches the CAD on, and gives a single assessment.

In the paper by Taylor, he showed that the concurrent reading may save time over the second reader. And that -- but however, the concurrent reading is still more time -- is more time consuming than reading without CAD. And the second reader paradigm may be more sensitive than concurrent reading.

Currently, regarding prospective clinical CTC CAD studies, currently we know of no prospective clinical studies in the literature investigating the effects of CTC

CAD.

Now I'm going to touch on some other issues specific to CTC CAD studies, including data acquisition, ground truth, or reference standards, and study endpoints.

As with all CADs, the CAD output is going to be dependent upon what you feed into it. And so, in this case, we have the patient, we have the CT scanner, we have -- the data gets processed, and it gets fed into the CAD, and then the CAD output comes out.

And all of these -- and each one of these steps has a number of different parameters associated with it, and each of these parameters may strongly, or may not, affect the CAD output. For example, patient parameters will include the bowel preparation method, insufflation, positioning of the patient.

CT parameters, of course, include slice, number of slices, slice thickness, the dose, the exposure, what reconstruction

algorithm was used for creating the image.

Then, there is also additional image processing that may fall in, including how stool was tagged and, as well as stool cleansing, digital, that is. And these will -- some of these points we refer to you, we're going to ask you questions about in question C3(a), regarding possible restrictions on these parameters and such.

The ground truth or reference standard. In other words, our -- is the CAD or the radiologist really marking true disease locations? How is this determined? What polyps or disease locations may have been missed during the study? There are a number of -- a number of different studies have used different ground truth or reference standards for determining this.

One method is to use optical colonoscopy. For example, the ACRIN Study, which -- whose numbers I cited earlier, uses optical colonoscopy as the gold standard. The

downside of this is that there have been reports that optical colonoscopy misses 11 percent or more lesions, and some of these lesions can be found by CTC CAD or, I'm sorry, CTC colonography.

I'm getting ahead of myself. The -- another possible method for ground truth is optical colonoscopy, and then expert review of the CT colonography data. Of course, the drawback with this method is that there may be variability among expert readers. And this variability should probably be accounted for.

Additionally, an additional method is optical colonoscopy while the optical -- while the colonoscopist has the report and findings from the CTC. This is often referred to as segmental unblinding,and can be found in Pickhardt, et al, for example.

Another, of course, possibility, is long-term follow-up. Whatever method is chosen for ground truth -- whichever of these methods is chosen for ground truth, matching

polyps between optical colonoscopy and any CTC scan can be rather problematic and can lead to biases and increases in variances in the study.

There are different methods, there are different matching methods and protocols for doing this, but I'm not going to go into these now.

Study endpoints, to wrap it up. What summary statistic should be used in a CTC CAD device study? Should we be concerned about sensitivity and specificity, or should we be primarily concerned with positive predictive value and negative predictive value?

Should we be requiring an ROC analysis or an FROC analysis? In any sort of ROC analysis, we would have to require that polyp or patient ratings/rankings be given by the clinician. However, this may require clinician training. This has been done in a couple of different papers, specifically,

Petrick and Taylor. Possibilities for doing this would be to use the C-RADS scale for patient ratings and/or using size or morphology of polyps to get a polyp rating.

And lastly, what, are there -- what should the relevant units of measure be? Should we be requesting the sensitivity by patient, or the sensitivity by polyp? Unlike in mammography, in colonography oftentimes patients will have several polyps or several lesions.

And thus, a by patient and by polyp analysis can be significantly different. Which polyps are relevant? Should we be wondering about the sensitivity of a CTC CAD device to all polyps? Should we only be worried about those which are more likely to be cancer such as adenomatous polyps, or should we be concerned only with polyps that are greater than a certain size?

And these will -- and these questions will be posed to you in questions

C3(a) and C4(a). Thank you.

CHAIRMAN GLASSMAN: Thank you very much. Does the Panel have any questions for Dr. Samuelson? No? Very -- no questions? That's very good. Thank you. It's now time for our second Open Public Hearing session. You are reminded that the same process, disclosure of any -- your affiliations, or anyone who is paying your way would be appreciated.

Again, it's five minutes, and our first speaker is Dr. Ron Summers, from the National Cancer Institute.

DR. SUMMERS: Thank you for the opportunity to speak today. I am a radiologist at the National Institute of Health, and I have 10 years of CAD experience, the majority of which is on the colon. My financial disclosure is shown here relating to iCAD and Viatronix.

The four important issues I would like to bring to your attention are patient

selection, image acquisition, performance benchmarks, and reading paradigm. And I'll explain why I think that's important.

It is my belief and my experience that most sensitivities of CAD systems reported in the literature are not accurate for estimating performance in the clinic. And I'm familiar with nearly all of the published literature for the colon in this area.

I believe this is due to several problems: selection bias, inadequate usage of common databases, and training and testing on the same data. The solution, in my view, is that the characteristics of the patient dataset must be clearly defined with detailed specifications.

And I want to highlight that I think this should be a consecutive series although a random selection from a consecutive series may be acceptable. Screening and diagnostic population should not be mixed, and demographic and recruitment info of the

patients, I believe, is mandatory. Also, a minimum dataset size needs to be determined.

I believe a separate test set and external validation are useful. External validation is where the tested CAD system is applied to a new patient population.

Regarding image acquisition and labeling, I would like to emphasize my opinion that broad labeling should not be considered acceptable, that the labeling should specify a preferred bowel prep and scanning parameters. The physician may decide, for a particular patient, to adjust these, but he or she should be aware that the CAD system may perform less well under such circumstances.

I believe proven protocol, such as the two I have listed, which have had high sensitivity in good trials, should be used. The prep could include, or should include, in my belief, cathartic bowel cleansing, oral contrast for fluid and stool tagging, and excellent colonic distension. The tagging

issue is a complex one, but I make that recommendation based on the two protocols I mentioned on the previous slide.

The imaging parameters, I think, should focus on detecting polyps 6 millimeters and larger. I've listed a number of things here. The only one I'll mention in detail is the third one because I have seen recommendations from important professional organizations that slice thicknesses as high as five millimeters may be acceptable, and I believe that not to be the case. So I think this is an important issue that needs to be considered.

Performance benchmarks should be laddered, based on the size of the polyp. I have given some thresholds for your consideration. The median false positive rate, I believe, should be roughly about 8 per patient. That should be expressed per patient, not per scan as I've seen in many publications.

Also, sensitivity should be reported based on all polyps, and polyps retrospectively visible on virtual colonoscopy. This is because some polyps found by the gold standard may not be visible on virtual, and I don't think it's fair to penalize the CAD for those.

Finally, reading paradigms, I have listened tentatively to the discussion this morning about reading paradigms. These also, apply for the colon. The first read is fast but has the lowest sensitivity in the limited studies that have been reported. The second read is slow but has the highest sensitivity. The concurrent read may be intermediate although there is very little published data on this at the present time.

My recommendation is the second read. As a physician, I think that's the appropriate one. And both the implementation and the labeling should discourage the first read by requiring the clinician to record the

pre-CAD read for auditing and accreditation. Otherwise, the physician may slip into a first read setting.

In conclusion, standardization and high benchmarks, I believe, will lead to CAD systems that are effective in the long-term, and help patients. Thank you very much.

CHAIRMAN GLASSMAN: Thank you. Next is Dr. Maha Sallam, from iCAD Medical.

DR. SALLAM: Thank you to the FDA, to the Panel, for the opportunity to speak. I just wanted to share with you a potential paradigm for approval of CAD for CTC devices.

CTC is emerging as a highly effective modality for the detection of polyps in the early identification of colorectal cancer. Some studies are already starting to show that it is possible to miss polyps on these exams. Even though they are highly sensitive, they are tedious to read, and it is still possible to miss polyps even when obvious.

iCAD, it is possible to design CAD systems that are able to do automated detection of these polyps in the readers. The CAD detection algorithm's technology consists of sophisticated software that has to be designed to distinguish between real polyps versus normal structures in the stool that may be in the colon.

Sensitivity and false marker rate are the standard benchmark numbers that characterize the standalone performance for a system. CAD systems are typically very sensitive but not specific enough. And so they are usually indicated for use as tools to assist readers and not as standalone readers.

And it is important to test the impact of CAD on the reader. The standalone performance is very relevant, but it's not necessarily indicative of that performance. For example, in a second reading scenario, a CAD system may be less sensitive than an independent radiologist, but in conjunction

with the radiologist, it may actually increase the detection rate by having an incremental addition to the detections that normally would be detected by the reader.

It is important to tie the testing methodology to the paradigm that is intended for the device or for the system. In a second read scenario, as I indicated, we can test the incremental findings that are found because of CAD, specifically, and it's possible to separate the individual radiologist reading from the incremental findings that are found by CAD.

In a first -- in a concurrent read, that may be a little bit more difficult because there is more interaction between the reader and the CAD system. And, as well as for first read.

Testing for a second read scenario as I indicated, can be done in two ways. To test the incremental findings because of CAD, but also to design reader studies that are --

that have a sequential design. Effectively, readers would read cases without CAD, followed immediately by reading with CAD, and then we're able to compare the performance of the two readings.

Again, with concurrent read, the interaction may be difficult. It may be difficult to separate the two readings and, hence, two separate reading sessions may be required if a reader study design is to be chosen as a method of evaluation.

For under a first read paradigm, that requires probably the most extensive testing in a standalone basis because we are completely relying on the CAD outcome to direct the radiologist's attention to areas within the image. So the standalone testing will have to be fairly extensive to be able to accommodate the different situations and conditions.

And again, it will not be enough to do just the standalone testing. It would need

to also be associated with the reader performance testing in addition to that.

So just in summary, I think in considering the appropriate testing, design testing methodologies, it is very important to distinguish between the different reading paradigms. It is very important to allow the vendors and the industry to specify the specific reading paradigm that they are proposing and recommending for the device.

There has been several studies that have been published already that reflect -- that are starting to reflect some findings based on the different study designs with the sequential -- with the separate reading sessions with the concurrent and sequential read and also with studies that already show first read as a possibility.

There is a lot of parallels. In the end, I would like to make a comment. There are a lot of parallels between mammography and CT colonography. The big

difference is that CT colonography is a little bit more wide open. And a CAD for mammography had been well-established for the last 10 years or so, and there are well-established methodologies for testing and approving those devices, and also support for that testing and approval in existing studies that are in the field right now.

With colonography, we have a different situation where the field is a little bit more open. More devices than not have not been -- have not gone through the approval process, and we have an opportunity to set things up based on the, you know, most current knowledge that we have in this area. And I thank you.

CHAIRMAN GLASSMAN: Thank you. Our third speaker is Dr. Gareth Beddoe, from Medicsight.

DR. BEDDOE: Hello there. I'm the Operations Director at Medicsight PLC, and I'm going to read this statement from the company.

Medicsight PLC appreciates this opportunity to provide comments to the Radiological Devices Panel for its consideration in connection with computer-aided detection and diagnosis devices.

Medicsight is a manufacturer of colon CAD devices, and we have over seven years of experience in this industry, and distribute devices in Europe, Canada, and Australia.

Medicsight welcomes clarification from the Panel and the Food and Drug Administration regarding the device types that are within the classification of CAD devices and clear guidance on the appropriate studies and statistical analyses required to support marketing clearance of these devices in the United States.

The FDA's Federal Register notice uses the term CAD to encompass both computer-aided detection and computer-assisted diagnosis, but a computer-aided detection

device presents different potential risks than a computer-assisted diagnosis device.

The stated indications for use for colon CAD devices for the detection of polyps required that radiologists review all images in the CT examination, and not only the images containing potential polyps identified by the software.

The device is thus intended to assist the physician interpreting the imaging studies. Medicsight requested the Panel and FDA recognize the distinction between the different types of CAD system and reflect this in the differential classification of computer-aided diagnosis devices and lower risk computer-aided detection devices.

The company also requests that the FDA limit the requirements for clinical data to those which satisfy the manufacturer's labeled intended use for a particular device. Manufacturers should not be required to produce evidence of device performance for

numerous off-label use scenarios because that is contrary to the statutory requirement that the FDA address potential off-label use scenarios through cautionary labeling.

And the following has been prepared by Professor Steve Halligan, from the University College Hospital of London, on behalf of Medicsight, PLC, and this is a discussion of the guidelines for statistical analysis for multi-reader, multi-case studies.

Medicsight asked that the FDA consider alternative approaches to multi-reader, multi-case receiver operator curve analysis for studies of CAD applied to CT colonography. CAD is employed in an ever increasing number of diagnostic scenarios, where CT colonography is relatively recent.

We agree that multiple readers and multiple cases are desirable and believe that studies of CAD for CT colonography should adopt the MRMC design. However, we believe that ROC analysis for such studies is

inappropriate and questionable. Such analysis has been applied to assist the detection of breast and lung lesions, but this stance can only apply to CT colonography if the diagnostic scenario is comparable.

In mammography, the clinical problem is one of detection and then lesion classification. A confidence score, upon which ROC analysis is predicated, is easy to assign, ranging from certainty the detected lesion is benign to certainty that it is malignant.

The situation for CT colonography is significantly different. Readers of CT colonography studies aim to detect polyps in the large bowel. There is no classification issue as the vast majority of polyps are benign. The clinical problem is thus one of detection and localization rather than classification.

Consequently, confidence scores regarding the likelihood of malignancy cannot

be applied meaningfully to colon polyps.

Regarding confidence scores for the presence of a lesion, a polyp first has to be detected by the reader. Having detected a polyp, the reader is then unlikely to assign a low probability to its presence. The result is that confidence scores for the presence of a polyp are positively skewed and highly non-normal which violates the fundamental assumption of ROC analysis.

A prior study found existing software could not fit close to 50 percent of readers due to degenerate data. Extensions to ROC methodology do not help since they are also predicated on confidence scores. I measured the number of false positive CAD prompts for a single 2D image, which is infeasible for CT colonographies, since the examinations are 3D volume.

Previous studies have also shown that readers identify few false positive polyps in patients with no polyps, i.e.,

specificity is high. The small number of false positive patients in polyps identified by readers means that the patients contributing to the ROC shape, and hence the area under the curve, are typically a small proportion of the study group.

This has an impact on power and confidence intervals, and the area under the curve is dominated by a curve extrapolated beyond the study data.

A prior study found no relationships between change in partial area under the curve and the more clinically relevant measure, that is, the increase in proportion of true positive and true negative patients correctly classified by readers, when CAD was used.

In ROC analysis, weighting of sensitivity and specificity is determined by software design. Indeed, weightings are arbitrary, and change contingent on the part of the ROC the data is obtained from, and so

are not influenced by clinical relevance.

However, it is clear that patients do not weigh sensitivity and specificity equally. A recent survey found that 56 percent of patients quoted diagnosis as the top priority, versus only 10 percent stating specificity.

At the present time, colonoscopy without prior imaging is the most commonly used method for colorectal cancer screening. Referrals for colonoscopy because of false positive detections by CAD are inevitable, but in reality these are patients who, in the absence of CT, would progress straight to colonoscopy in any event.

Numerous issues make ROC analysis inappropriate for CT colonography studies. These issues are conceptual. Confidence scores cannot be meaningfully assigned. Statistical data are highly non-normal. Practical existing software cannot fit curves, and ethical, patient's personal weigh-ins are

ignored.

We ask the Panel and FDA to recognize that alternative to analysis plans for MRMC studies are appropriate, and the FDA can carefully consider alternatives to ROC analysis when offered by statistical groups collaborating with industry partners. Thank you.

CHAIRMAN GLASSMAN: Thank you. At this time, the Chair has decided to allow Dr. Akira Hasegawa an additional five minutes to complete his presentation from this morning from Fujifilm Medical Systems.

DR. HASEGAWA: I have five minutes?

CHAIRMAN GLASSMAN: That's right, five minutes only.

DR. HASEGAWA: Just, I wanted to make sure. So I think I almost finished this one. So I think I probably read the Type 3 CAD, it's for concurrent read. Risk is, it may cause users' satisfaction of search, and it may affect users searching negatively.

In the previous slides, we explained how each type of CAD affects the standard reading procedure. As a result, we developed a decision tree for assessing CAD risk.

First question is, human makes final diagnosis, or no? If no, then it is not CAD. And it is automated detection, or diagnosis device.

And, if yes, the next question is, all images are reviewed by human. If no, then it is not CAD, either. And it is computerized screening device. FDA refers to this as a first reader mode on page 13.

Because both these device makes some kind of diagnosis without radiologist, the risk level is high. But because these are not CAD, so, in the following presentation, one focused on this device.

And go back to the question. If yes, then the next question is the device is used for diagnosis. If no, the next question

is, the device may affect humans searching process. If no, this is Type 1 CAD for optional second read.

The risk of this device -- the risk of this type of CAD is low. If yes, it is Type 3 CAD for concurrent read. Then, compared to Type 1, this CAD may affect human searching process, so its risk level is a little bit higher than Type 1. So the risk of this type of CAD is low in immediate risk.

Then go back to the level that this -- the devices for diagnosis. If yes, the next question is the device provides classification, such as benign and malignant. If yes, it is a Type 2 CAD for interpretation, and the FDA refers to this as computer-aided diagnosis.

This type of CAD suggests classification to radiologists, so the risk level is higher than Type 1 and Type 3. But still, the radiologist makes the final decision, or diagnosis. So compared to

automated detection, or diagnosis device, the risk is much lower. So the risk of this type of CAD is moderate risk.

If no, such device do no more than computerize measurement or segmentation that can be reviewed by -- reviewed and edited by readers. Obviously, computerized measurement is kind of replacement of past ruler, and the segmentation is an accumulation of measurement. So the risk of such device is much lower.

Summary: all CAD are not the same. Different CAD has different indication for use. Different CADs has different risk, or risk factor. Based on this, we proposed an example of this decision tree for assessing risk of CAD.

As explained in this presentation, three different CADs are totally different device although they are all called CAD. They have different eye view, and different risk level. I read the FDA's document. I think

it's well-written. But I saw, in many places, these three different of CAD are mixed together and discussed together.

I think we have to re-distinguish these three type of CAD. Thank you very much.

CHAIRMAN GLASSMAN: Thank you. Are there any questions from the Panel of any of our speakers for this afternoon? If not, we will -- I'm sorry, Dr. Berry?

DR. BERRY: So I would like to ask Dr. Samuelson just to clarify, in view of the presentations we just heard, the -- when you talk about sensitivity and specificity, you ask a question about polyps and adenomatous polyps. One of the speakers talked about cancer. When you say sensitivity and specificity, what are you talking about? Is it just polyps, some category of polyps, or are you talking about cancer?

DR. SAMUELSON: Well, that's one of the questions we're posing to you is, should it be cancer? Should it be polyps? In

general, when making a diagnosis on a CT colonography or even an optical colonoscopy, it's difficult, or almost impossible, for the radiologist or the colonoscopist to tell the difference by visual inspection as to whether that polyp is actually malignan, or not.

And thus, the question then becomes, well of course, cancer is what we're shooting for, and it would be nice to say, we only care about cancer. But at the level of the data that we're receiving, optic which is the CT colonography or the optical inspection from ROC, one can't really tell the difference, at that point. And so it's not until it goes to biopsy.

DR. BERRY: So you could still talk about sensitivity and specificity for picking up polyps, and the Medicsight speaker said ROCs aren't relevant. But I think it's your view that they are relevant if you are focusing on just polyps, for example. And they could be relevant for cancer, as well, if