NATIONAL QUALITY FORUM

Moderator: Cancer -
May 3, 2016
12:00 p.m. ET

OPERATOR: This is Conference #: 86671994.

Operator*:* Welcome everyone. The Webcast is about to begin. Please note, today's call is being recorded. Please stand by.

Shaconna Gorham: Good afternoon and welcome to the Cancer Care Standing Committee First Workgroup Call. My name is Shaconna Gorham and I am the Senior Project Manager for this project. And Amber?

Amber Sterling: My name is Amber Sterling. I'm the Project Manager for this project.

Melissa Mariñelarena: And this Melissa Mariñelarena. I'm a Senior Director.

Shaconna Gorham: And Kaitlynn Robinson-Ector is our Project Analyst. And she's unable to join us today, but we have Donna Herring. And she is also a Project Analyst to NQF, and she will be helping us today. So, welcome.

I would like to just do a roll call to see if our committee members have joined us.

Jennifer Harvey?

Jennifer Harvey: Yes. Yes, I'm here.

Shaconna Gorham: Shelley Fuld Nasso?

Shelley Fuld Nasso: Yes, I'm here.

Shaconna Gorham: Jennifer Carney?

Jennifer Carney: Yes, I'm here.

Shaconna Gorham: Joseph Laver? David Cella? And Jennifer Malin?

OK. If you have logged on to the web, we also need you to dial in.

With that being said, we would also like to welcome our developers. Do we have a representative from the American College of Radiology?

Judy Burleson: Hi, this is Judy Burleson, Alicia Blakey from ACR. And I believe Dr. David Seidenwurm is on also.

Shaconna Gorham: OK. The American Society of Hematology?

Sam Tierney: Hi. This is Sam Tierney. I'm with the PCPI. We are supporting the American Society of Hematology and the submission of these measures. And I believe we also have a physician on the phone, Dr. (Able).

Shaconna Gorham: All right, (who)?

(Dr. Able): Yes, I'm on the phone.

Shaconna Gorham: OK. Perfect. Centers for Medicaid and Medicare?

Christine Holland: Hi. This is Christine Holland at Mathematica Policy Research. And we'll be supporting CMS in the submission of one of the measures.

Shaconna Gorham: OK. Perfect. So, welcome everyone. The purpose of today's call is to allow the Standing Committee members to have preliminary discussions about measures that will be evaluated during the in-person meeting.

So, many of you all are new to the NQF process. So, this is really our opportunity to ask questions about the criteria, expectations of you as committee members. And please remember that the developers are on the line

to answer questions or to clarify anything that may have given you pause as you reviewed your measure worksheets.

NQF staff will introduce the measures and moderate a bit. So essentially, we'll be acting as the co-chairs, and then we would turn the conversation over to the lead discussants.

As the discussants – once the measure discussion unfolds, we will screen-share measure worksheets. So (Donna), can we put this first, the worksheet up, just so that we ensure that everyone is on the same page.

(Donna): Are we doing 508?

Shaconna Gorham: Correct. Yes.

So, this is the measure worksheet. And if we scroll down, I just want to show you the section for the committee comments. Keep going. So, right there. So, depending on your computer, the area for committee comments would either be pink or peachy. But the discussants will summarize the information presented by the developers, as well as the comments submitted by each Standing Committee members. And those that were in the pink or the peachy section is where the Standing Committee comments would be. So, we can scroll back up.

So, our first two measures today have been submitted by the American College of Radiology, and they are 0508 and 0509. And our discussants are Jennifer Harvey and Shelley Fuld Nasso. And so, the first measure is 0508, Diagnostic Imaging: Inappropriate Use of " Probably Benign" Assessment Category in Screening Mammograms.

And our first criteria is importance to measure and report. And we are going to look at the evidence and the opportunity for improvements and gap. And so, remember that this is a maintenance measure, so there is the increased emphasis on gap. So, I would turn it over to Jennifer and Shelley.

Jennifer Harvey: OK. So, this is my first time.

Shaconna Gorham: OK.

Jennifer Harvey: And, of course, I get the first case.

Shaconna Gorham: It's OK. So, we just want you to summarize the information that the developer gave, as well as the comments from the committee.

Jennifer Harvey: Sure. So, Probably Benign is (inaudible) BI-RADS 3, which basically there are quite specific indications to give a BI-RADS 3 recommendation which is short interval follow-up. We never – we should never give it from screening because diagnostic evaluation with diagnostic mammography and ultrasound is known from the literature to identified cancers as well as benign lesions. And so, those are both inappropriate for short interval follow-up, so it reduces the number of BI-RADS 3 things. So, what this measure identifies is that is to monitor how often BI-RADS 3 or Probably Benign is given from screening, which should be never.

So, that's my summary, so, which you want.

Shaconna Gorham: Perfect. And was that Jennifer speaking?

Jennifer Harvey: Yes.

Shaconna Gorham: OK. Shelley, did you have anything to add?

Shelley Fuld Nasso: Well, I guess my question is just if you should never give it, and there're really not any exemptions, why is it even an option? I mean, it just seems like it shouldn't be – you shouldn't be allowed, I guess, maybe it's like that.

(Crosstalk)

Jennifer Harvey: Well, and I don't know why.

Shelley Fuld Nasso: … could prevent it.

Jennifer Harvey: I don't know why but people do use it. You know, essentially from screening, your choices should be BI-RADS 0, 1 or 2, which is basically recall or no recall. But people do use the other categories which is inappropriate. And,

you know, whatever they dictate is what is going to be entered out. I mean, obviously, there's going to occasionally be an error. But really, you're right. It should be 0, but it's not unfortunately.

Shaconna Gorham: OK. Do we have comments or questions from other committee members as well?

OK. OK. Jennifer and Shelley, I'll turn it back over to you.

Jennifer Harvey: OK. What would you like now?

Shaconna Gorham: So, unless you have questions for the developers, we can move on to opportunity for improvement in gap.

Jennifer Harvey: My only question would be, there was somebody on from CMS, right?

Shaconna Gorham: No, there wasn't anyone on from CMS. There's somebody on from Mathematica representing a different measure. But there is – the developer is on the line, which is ACR.

Jennifer Harvey: OK. Now, I do think this an important measure. You know, as a practicing radiologist, this is one that I know well. So, I think that it's been very successful in highlighting the importance of how to manage and use BI-RADS 3 lesions. So, I have no concerns with renewing this.

Shelley Fuld Nasso: This is Shelley. I have two questions. One is it's clearly the performance on this measure is very high now that you can see in the data that it has increased over time, and probably because of this measure. So, at what point do you – is it not necessary to measure anymore, because it is standard practice? And then – well, I'll ask that questions first and then I have another question too.

Jennifer Harvey: I know, I think that's a really good point. It has performed very well and has done a lot to increase the awareness, as I said earlier. But I'm trying to find where I look at the numbers. You know, if it's 99 percent, that still means that a lot of women are getting a BI-RADS 3 from screening. Does that make sense? Like in my practice, we read 30,000 screens a year. If 1 percent of

those is getting BI-RADS 3, that's too many. So, it should be pretty close to zero.

Shelley Fuld Nasso: It makes sense. And then I think the fact in this one that it's not been a 99 percent every year. I mean, it's only gotten to that point. So, it does make sense to continue measuring, but I know that's sort of a philosophical question overall when the measure is so high that it's not – there's not a big gap that needs to made up.

Jennifer Harvey: No, I think that's a really good point. So how – this would not be renewed again for three years or six years?

Melissa Mariñelarena: This is Melissa. If it gets re-endorsed under maintenance, it's endorsed for another three years.

Jennifer Harvey: OK. You know, I think I would like to see how that looks at another three years. If that is maintained at the 2014 level, then it should be discontinued because it's done its job.

Shaconna Gorham: OK.

Shelley Fuld Nasso: This is Shelley. I have another question. On the under threats to validity or the guidance for the validity algorithm, it says insufficient preliminary rating insufficient. And I'm not exactly sure what does that mean for our analysis or how do we need to consider that.

Shaconna Gorham: So, Shelley, are you speaking of evidence and not validity? We haven't gotten into validity yet. So, and we do have a preliminary rating for evidence as insufficient.

Shelley Fuld Nasso: OK. Sorry. No, I was talking about – I've gone down to validity. I didn't realize we're still on evidence. I thought we were just sort of talking generally about it. But that's fine.

Shaconna Gorham: So, we want to …

(Crosstalk)

Shaconna Gorham: We do want to follow the criteria. So, we started with evidence and then we moved to performance gap. But if we want to go back to evidence, we can, why we're still, you know, here for this criterion.

Jennifer Harvey: No. So, I, you know, I am very new to this. Why would this be considered insufficient evidence?

Melissa Mariñelarena: This is Melissa. This was only based on the algorithm that we use and that was because the level of evidence was not graded. And when you walk through the algorithm, that's how we ended up. But then, you'll also see there's the option here for exceptions to the evidence. So, you know, as a professional, if you know that there is, you know, there is no empiric evidence to support this, and that happens sometimes, right? There's not going to be any RCPs for something like this.

Jennifer Harvey: Right.

Melissa Mariñelarena: So, you can walk through the algorithm and then determine that an exception is justified for a measure like this. So, that was why the option is there. It's just that, you know, there's no RCPs, the systematic review. It just doesn't fit with our criteria under evidence. That was (also).

Jennifer Harvey: So, not even with the paper by the Breast Cancer Surveillance Consortium is showing, you know, from – it's in one of the appendix. It's referenced anyway. Sorry.

I'm going to find it. Reference 7, let's see. Can I give you a page number? I don't see a page number. The study by Karla Kerlikowske, even before this became a measure, demonstrated that, you know, by using over a million screening mammograms, that women were undergoing inappropriate short-term follow up when used from screening.

Shelley Fuld Nasso: Is that one of additional studies?

Jennifer Harvey: Yes.

Shelley Fuld Nasso: Because they do refer to be additional studies. It's just …

Jennifer Harvey: Yes.

Shelley Fuld Nasso: … when you go through the algorithm, and that's fine. You know, the committee, when you discuss this, that's fine.

Jennifer Harvey: Yes.

Shelley Fuld Nasso: You know, because we just tried the one guideline which is not graded, and then all the additional guidelines on breast cancer screening would have to do more with the schedule and the recommendations for breast cancer, which don't exactly align with the focus of this measure, which focuses more on the Probably Benign category. You know, and then we just discussed – we don't go with the detail about the eight studies, but you could talk about that. And again, because this is not really – you know, there's not going to be necessarily empirical evidence for something like this on a, you know, what you should not be diagnosing.

Jennifer Harvey: Yes.

Shelley Fuld Nasso: So then, at the meeting, we could talk about that. And the way something works for a vote and this is active because its maintenance. It was already recommended before, we don't have to vote through it even though …

Jennifer Harvey: OK.

Shelley Fuld Nasso: … there were updates to the evidence, but it's not really much different. I think it's just – if anything, we got more evidence than we had back in 2008. So the committee can say, "You know what, we're fine with what the committee voted back in the original recommendation back in 2008. We don't feel a need to vote again."

Jennifer Harvey: OK. I see.

Shelley Fuld Nasso: Or we could go to the former vote where it actually has to fail on evidence. We have to get a vote of insufficient, and then we go through the algorithm and then decide that it fits under this exception to the evidence.

Jennifer Harvey: OK.

Shelley Fuld Nasso: And then we would write up the rationale for that. But, you know, we can discuss that at the meeting, then we can refer to the study that you're talking about, or we can even write that up just because the evidence – there is more evidence than it was back in 2008. But it's just the algorithm.

Jennifer Harvey: OK. I see. All right.

Shelley Fuld Nasso: Yes.

Shaconna Gorham: OK. So, if you ladies are satisfied with the conversation for evidence and performance gap and we have no other comments from standing committee members, we can go to the next criterion which is reliability.

Jennifer Harvey: I think the reliability is high because of the way that this is evaluated using the billing data.

Shaconna Gorham: OK, and that was Jennifer speaking. Shelley? Did you have anything to add, Shelley?

Shelley Fuld Nasso: Oh sorry, I was talking with the phone on mute. I think it looks like it's really straightforward and the reliability is very high.

Shaconna Gorham: OK. Additional comments from other Standing Committee members? No?

OK. I'll turn it back over to Shelley and Jennifer for conversation about validity.

Shelley Fuld Nasso: Well, this is Shelley. This is where I have the question about insufficient, because I guess I don't understand the algorithm and why the algorithm says this is insufficient, especially when it's been endorsed and it's in use.

Melissa Mariñelarena: Right. So, it's because the – what we require now is not the same as what it was even four years ago. And the last time this was looked at was in 2008. So when you look at the algorithm for validity, the first thing that it asked for is, "Were threats to validity assessed?" So, if you look at – there were no

exclusions in the measure, so that does not apply. We do ask you, do you think there should be exclusions, but that doesn't imply.

The next thing that we asked for are meaningful differences, just for them to be addressed. The developer could come back and just give us something. And it doesn't always – we don't always get statistical analysis in here, but just to address that. The next section is comparability of data sources. That usually doesn't apply unless there is a measure that has different sets of specifications.

And missing data, we – again, there was nothing provided here and we don't always get a statistical analysis. But if it doesn't apply because the measures is claims, OK, that just somehow that it was – that these threats to validity were assessed because the way the algorithm works, that's the first thing that we have to address. And if they're not assessed at all, then it goes to rate as insufficient. But again, that just means there's not enough information here for us to get a rating on validity.

Shaconna Gorham: OK.

Jennifer Carney: And this is Jennifer Carney. I'm just wondering, but since the use of the face validity as a testing method, it still doesn't rate as moderate or low?

Melissa Mariñelarena: Let me look at here. It does. So the face validity is validity testing, which testing – validity testing is different from the threats to validity. They have to be addressed separately.

Jennifer Carney: I see. OK.

Shaconna Gorham: I see.

Melissa Mariñelarena: And for our face validity is the minimum thresholds for validity that we asked for. And (duly) very specific what we ask for, and it's here in Italics, where, you know, we require the face validity testing results indicate that the measure aspects, but it can be used to distinguish good form of quality. And I think they offer – there were different responses here. And the one that it's

hyperlinked too is the one response that is most closely linked to what we look for in face validity.

Shelley Fuld Nasso: OK. So, this is Shelley. So basically, the reason for that as insufficient based on the algorithm is because there wasn't information provided about meaningful differences and missing – and threats to validity.

So can – are the – could the developers address that? I mean is that – is there something we need to be concerned about here or? I guess, I just – I don't know what the criteria are for us to sort of bypass the algorithm if data is not provided, that it needs to be provided to get the right rating.

Melissa Mariñelarena: Right. And I mean that's a question that you just, you know, post the developers in there, you know, if they like to respond as well. There're also – are there questions for the committee and their meaningful difference? We also say, you know, given the data provided in 1B, which is their performance data, we ask you, you know, as a committee, can you make a judgment on, you know, based on that data, can you identify meaningful differences about quality across physicians?

The information that we asked under meaningful difference is very specific. And we ask that it's different from performance data, sometimes we get just performance data, but we ask you. You know, so look at this information, can you tell, can you identify meaningful differences? And, you know, you can discuss that if you want to take a look at that, but that's a great question to the developers, and I don't know with anybody from ACR wants to respond?

David Seidenwurm: May I? This is David Seidenwurm on the phone.

Melissa Mariñelarena: Yes, please.

David Seidenwurm: Great. By the way, thank you very much for considering this measure. We at the college think this is extremely important for a number of reasons. The principal reason that we think that this is an important measure is that it leads to appropriate care in several ways. One way is the way that Dr. Harvey explained, which is that women are cared for correctly when they have

findings that are – look like they might be probably benign on a screening mammogram but haven't been fully evaluated.

The second reason and equally important is that this measure, even though it appears to be close to being tapped out is actually sufficiently – there's a sufficient gap in care that could compromise the integrity of the hospital compare mammography recall metric, because as the group may or may not be familiar with that measurement, let me elaborate a bit.

The hospital compare metric calculates and administratively drive proxy from mammography recall rate. And the national average for that proxy recall rate in the Medicare population is approximately 8 percent. So, a 1.6 percent or 2 percent non-compliance with this measure can throw off the hospital compare recall measure by a quarter, and can result in miscategorization of substantial numbers of physicians. So, this could have a lot of impact, a lot of practices, I think it's done at the practice level.

So this metric that we're discussing today is actually extremely important and we need to think of its gap in care as closer to 20 percent or 25 percent rather than between 1 percent and 2 percent for the reasons that I said.

As with respect to the threats to validity, there are heterogeneous practice patterns, and I'm not quite sure why we were unable to submit data on that. There are physicians that report up to 2 percent of their – 2 percent or 3 percent of their mammography population or screening mammography population in this way, and others that the majority at zero or near zero. I'm wondering if, perhaps, we can dig up some data from the registry and submit that when the measure is considered, if that would be in order.

So anyway, the principal reasons to support this measure, even though it may appear to be close to being tapped out or that that's small, what seems like a small percentage is actually a large percentage of another measure that's used to compare practices in this important area. And anyway, thank you for considering it. And if there's any other help we can be, let us know.

Shaconna Gorham: So, I just wanted to make a note. We definitely, here at NQF, will accept additional information that you have. We can take that information and prepare it and then, you know, add it to the measure worksheet in preparation for the in-person meeting.

David Seidenwurm: Thank you very much.

Shaconna Gorham: Yes. OK.

Shelley Fuld Nasso: This is Shelley. I thought that was a helpful explanation too because I was wondering if the, you know, if you can see the variation by a physician or by practice. So, that's helpful. Thank you.

Jennifer Harvey: It's a really good point.

Shaconna Gorham: All right. All right ladies, Jennifer and Shelley, I will turn it back over to you for discussion on feasibility.

Jennifer Harvey: The feasibility is high because the data is already (inaudible) electronically, so it's easy to access.

Shelley Fuld Nasso: Yes, I have nothing to add.

Shaconna Gorham: OK. All right, comments from other Standing Committee members?

All right, we'll move on to use and usability. Again, this is a maintenance measure, so there is the increased emphasis on usability and use.

Shelley and Jennifer?

Jennifer Harvey: I agree with moderate for the reasons we've talked about, yes.

Shaconna Gorham: OK.

Shelley Fuld Nasso: And this is Shelley. I don't really know how to assess that. I mean, it seems like that it's obviously been – well, so the performance has improved over the years. So, I would assume that somebody is using it to help – you know, that that's helping to drive that. Maybe that's not the reason, but it

seems like it's helping to improve the performance. So, even though it's not publicly reported, it seems like it has some value.

Jennifer Harvey: Yes, I would agree.

Shaconna Gorham: OK. So hopefully, the discussion on the measure was helpful. The only difference in this discussion and in the actual in-person meeting is we will actually take a vote after each criterion. But as lead discussants, you will still lead the measure as you did today, and then we'll actually vote.

So, we'll move on to the next measure, which is 0509. And again, our lead discussants are Jennifer Harvey and Shelley Fuld Nasso. And the title is Diagnostic Imaging: Reminder System for Screening Mammograms. And again, our developer is American College of Radiology. So, I will turn it over to Shelley and Jennifer for evidence.

Jennifer Harvey: All right. Would you want me to go, Shelley? I'm happy to do so.

Shelley Fuld Nasso: Yes, that would be great.

Jennifer Harvey: So again, you know, I'm a practicing radiologist. I do breast imaging. So, I think this measure is becoming even more important actually. So, you know, in the past, we have typically recommended screening every year and sending out reminder letters.

So this measure basically evaluates how often practices are sending reminder letters for women to get their next screening mammogram. And in, you know, the world of getting a mammogram every year, that seems pretty straightforward. But in practices that may be recommending mammography every two years, the reminder letter really is crucial because the difference between screening every two years and every three years is really significant if you – one of the data references that they provide is screening data from the U.K., where they do three-year screenings. And the interval cancer rate, which is, you know, a palpable cancer appearing between screens is very high in that year, between the second and third year. So it's really imperative that women show up by two years for their screening. And we know that reminder

systems work pretty well for getting women in for screening. So, that's sort of my summary.

Shaconna Gorham: Thank you. Shelley, anything to add? All right.

Shelley Fund Nasso: Oh, I'm sorry. I was talking on mute again. And I thought Jennifer did a great job explaining it. I just thought there was no good evidence presented that shows how reminders help with compliance with screening recommendations. So, it just seems very worthwhile and the evidence is good.

Shaconna Gorham: OK, other comments from committee members? All right, we'll move on to performance gap.

Shelley or Jennifer?

Shelley Fuld Nasso: This is Shelley. There definitely still are gaps although the performance is improving, again, over the years. I was a little bit confused about, and now, I'm worried that I'm mixing up which measures I'm talking about. That there was this – there was a dip in 2013 and there were some explanation about why that was, but I didn't understand that. And then it went back up, but I didn't understand that dip. Am I getting – am I confusing which measure I'm thinking of?

Jennifer Harvey: I think so, yes.

(Crosstalk)

Jennifer Harvey: … was the dip, yes.

Shelley Fuld Nasso: OK. Maybe I'm thinking it's the next one. Sorry. OK, never mind.

Jennifer Harvey: It's OK. Again, I think although this is pretty good, I am concerned that as guidelines are shifting, that this could get worse. So, I think it could be really helpful to monitor.

Shaconna Gorham: OK, additional comments?

All right, we can move on to reliability.

Jennifer Harvey: I thought that it's quite high for the same reason as the last one. It's electronic, so it's pretty easy to do.

Shelley Fuld Nasso: Yes, I would agree.

Shaconna Gorham: So, we want to look at reliability, the specifications and reliability testing. I think you're referring to feasibility now.

Shelley Fuld Nasso: Yes, you're right. OK.

Shaconna Gorham: So, do we have any comments on the reliability testing of the measure?

Jennifer Harvey: I mean, it looks like, you know, when they looked at practices and try to reproduce some, that it – I mean, they showed 100 percent agreement, so it looks like this is very reproducible.

Shaconna Gorham: OK, other comments?

All right, we can move on to validity.

Jennifer Harvey: OK. So this is one that have same issue of insufficient because the threats to validity were not identified, so I guess I would ask – maybe ask the developers again to address that and maybe if there's additional information you'll provide, that would be great, but if you have anything to add, I think that's great.

Shaconna Gorham: Sure. Did the developers want to address the comment?

David Seidenwurm: Sure, David Seidenwurm here again, and from the American College of Radiology. Thanks again for considering this metric. You know, we at the college, think it's extremely important. As what's said earlier, with the rapidly changing guideline environment and especially with the lengthening of intervals between screening as recommended, for example, by the preventive services task force and others. We think that having systematic reminders is extremely important in order to preserve the effectiveness of mammography

while minimizing the harms of frequent screening. So again, that's the rationale behind the measure.

With respect to threats to validity, the heterogeneities in practice with this measure seemed to be binary. Either there is a systematic attempt to manage the recall of women for their subsequent screens. I shouldn't use the term recall. So, remind women of their subsequent screens, you know, at whatever the agreed upon screening interval is for that particular practice. Either the practices for doing that or they're not, we think that this should be near 100 percent because the efficacy of screening is achieved somewhat by the initial or prevalent screen, but principally through the sequence of screens that occur after that. So it's essentially binary. And I'm not sure where we would have the data except in the threat to validity, except that there's compliance or non-compliance. Anyway, thank you very much for considering the measure.

Melissa Mariñelarena: Hi, this is Melissa. I have a question. It looks like there was the exclusion medical reason documentation was added in 2014. Is that correct?

David Seidenwurm: I guess so, yes.

Melissa Mariñelarena: And so, do you have an analysis of the exclusion, like how many times it's been used, how many times per clinician, anything like that since it was added?

David Seidenwurm: You know, I don't have that in front of me. What we – we can get that for the meeting, I think. The reason that we wanted to have exclusions, generally, we would prefer to have no exclusions, but we wanted to make sure that women, for example, with less than 10 years are reasonable life expectancy and other reasons to not pursue for their mammography would not be receiving reminder letters. So that was, you know, one of the reasons we put that in there. I'm not sure if we have hard data on the frequency with which that's used.

Judy Burleson: Hi. This is Judy Burleson at ACR. We would have to get that from – to have good numbers on that, we'd have to get more data from claims reported, the measure being reported by claims from CMS. I'm not sure if that was in the

data that they gave us that we did most of the reliability review on. We might be able to get a bit from practice or individuals that have reported through our registry, but it wouldn't be very – they'd just be small numbers. So, I don't know how valid that would be on the exception.

Melissa Mariñelarena: OK.

Judy Burleson: But we'll see what we can get.

Melissa Mariñelarena: OK. Thank you.

Shaconna Gorham: Do we have anymore comments from committee members? And I neglected to say in the beginning of the call. Although we have a lead discussants designated, if you are a committee member and you have comments on these measures, even if you were not designated as a lead discussant, please feel free to add your comments. We welcome them.

So with that said, do we have any comments from committee members even though not assigned to this workgroup?

OK. So, I'll turn it back over to Shelley and Jennifer for discussion on feasibility.

Jennifer Harvey: Feasibility is high. The data, that's pretty easy to get.

Shelley Fuld Nasso: Yes, I don't have anything to add to that.

Shaconna Gorham: OK.

Shelley Fuld Nasso: It seems like pretty (straightforward).

Shaconna Gorham: OK. We can move on to usability and use.

Jennifer Harvey: So, this has been used as a PQRS. Even though it's not publicly reported, it is a known quality metric. And it does look like it has been used in accountability programs. So, it seems like it's got reasonable usability.

Shaconna Gorham: OK, additional comments?

Judy Burleson: Would a developer comment be appropriate at this time?

Shaconna Gorham: Sure.

Judy Burleson: Thank you. So, this measure by itself is not publicly reported. But the mammography uptake rate is reported in numerous programs. So, this does impact a publicly reported measure and that attests to its importance.

Jennifer Harvey: That's a good point. Yes. Oh, you mean because that their relationship to the breast cancer measure. Is that what you mean?

Judy Burleson: That's correct. Yes.

Jennifer Harvey: OK, got it. That makes sense.

Shaconna Gorham: All right, so that concludes our discussion of 058 and 0509. We will move on to our next set of measures. And these measures have been developed by the American Society of Hematology. And our first measure is 0377. And our lead discussant is Jennifer Carney. I do not believe that Joseph Laver has joined this call.

So Jennifer, I will turn it over to you for discussion on evidence first.

Jennifer Carney: OK. Thank you. And I'd like to just say like Jennifer Harvey, I'm Jennifer Carney with a C. But I also – you know, I do all types of hematology and oncology. And so, it's my first time on this. So, just please help me out if I need some help.

Shaconna Gorham: Of course.

Jennifer Carney: I'm excited about this measure because, I mean, it really is, you know, the first form of pursuing medicine which is such a big emphasis in oncology now for nearly diagnosed acute leukemia or myelodysplasia. Cytogenetic testing remains critical for diagnosis, prognosis and our therapeutic options.

So, this maintenance measure was originally endorsed in 2008, last endorsed in 2012. And so, this is a maintenance measure. And so, I'm supposed to talk

about, you know, about the level of evidence which remains based on the NCCN guidelines is graded to a lower level of evidence. But there is uniform consensus that this intervention is appropriate because this is basically the baseline and one of the most important test we utilize for our treatment.

So, that's my introduction. After review of the current literature, in addition, I couldn't find anything either with a higher level of evidence. I would just like to comment because there were some comments on the last review in 2012 about the level of evidence for newly diagnosed acute leukemia. But I don't think that those type of studies would exist in this day and age, as commented earlier that there are some role for empirical evidence in this setting now. Primarily studies were – are looking more at (longitudinal numerals) or other cytogenetic risks for stratification of treatments. So, I think that there is no new evidence. And then, I guess, the next – so we would rate that as low.

Is that good?

Shaconna Gorham: Yes, that's perfect. I would just like to add that this is a maintenance measure. And the developer attest that the underlining evidence have not changed since the last NQF endorsement. So, the committee have – has the option of voting or not voting on this subcriterion.

So, if there additional comments from other Standing Committee members, we would accept them now.

Jennifer Carney: This is Jennifer Carney again. I would just say, I guess in the last review, one thing that was brought up was whether or not, you know, in some ways, because of the rapidity of this field, the cytogenetic testing alone is, you know, an appropriate measure. And whether or not, I think the prior panel discussed whether or not adding FISH or other molecular testing was appropriate to consider. But I didn't – wasn't sure that I saw from any comment, response from the developers.

Shaconna Gorham: Would the developers like to comment?

Sam Tierney: This is Sam Tierney with the PCPI. I just wonder, Dr. (Able), do you have any perspective on adding those additional molecular testing? I believe, you

know, the measure states some of the guidelines, and I think consistent with that. But I don't know, Dr. (Able), if you have any perspective on that.

(Dr. Able): Sure. I think that is – adding molecular genetic testing is something that is coming. But I don't know if the data that we have support doing it yet. Right now, the WHO guidelines, the WHO categories for putting patients into different buckets in terms of risk and morphology do incorporate some cytogenetics but don't have a lot of molecular genetic testing. So, I don't think that the evidence is there yet to add that in. But it is something that I think we need to think about in the future with this measure.

And also, I think helps support this measure because it's, you know, other further testing to be done bone marrow. So, this measure sort of becomes don't do a bone marrow without additional testing on that marrow which is cytogenetics in this case. And I think it is a good setup for adding molecular.

I think FISH is, definitely, there is an evidence for FISH at this point. And I think molecular genetic testing, I don't think that there's enough evidence on what to do with that information treatment-wise for all kind of MDS and AML that it should be required. But the cytogenetics, its part of the very definition of some of the subtypes that you have the cytogenetic types, so I think that continues to be supported and is important.

Shaconna Gorham: Thank you. So Jennifer, I'll turn it back over to you. We can move on to performance gap?

Jennifer Carney: Yes. So in 2014, there was a wide distribution of performance rates of a marrow being performed for the cytogenetics. There was also from the PQRS data rates, the performance rates are in the, you know, high 80s to the lower 90s. You know, this does suggest that quality issue or problem – the range of physician performance from 0.26 to 1 suggest meaningful variation across physician's performance to an opportunity for improvement from the data. And disparities data is not available.

Shaconna Gorham: All right. Thank you. Additional comments from Standing Committee members?

OK. We can move on to reliability.

Jennifer Carney: So reliability testing …

(Off-Mic)

I feel very lonely without my partner.

Shaconna Gorham: You're doing great.

(Off-Mic)

Female: We'll help you.

Jennifer Carney: So, this is 2A, 2B and 2D. So, I said that the data elements are clearly defined, that calculation algorithm is clear. Oh, am I looking at the right thing? Hold on. I'm sorry. I didn't actually bring my hard copy which I've written.

Shaconna Gorham: No, you're fine. You're going over reliability specifications. You're OK.

Jennifer Carney: OK. Yes. I mentioned about that the measure can be implemented consistently. That it's understandable that the sample size is small for basically a more rare disease. And that the specifications are consistent with evidence. Oh yes, and here, inter-rater reliability overall is 98.3 percent. That was the prior with this test sample of 67 physicians. They used signal-to-noise for their analysis reliability score and it was 0.82. So, that's the level that's acceptable for reliability.

Shaconna Gorham: OK. Additional comments from Standing Committee members?

All right, we can move on to validity.

Jennifer Carney: So in this measure, maintenance measure, they used face validity, which is a minimal but acceptable way to test your validity. There was substantial agreement between the ASH experts at 94 percent. And there were no threats to validity, potential for bias. And so, I rated it moderate.

I made one comment about the overall exception rate with 1.2, which is relatively low. Exclusion analysis is not a threat to validity.

Shaconna Gorham: All right. Additional comments from Standing Committee members or questions?

All right, feasibility.

Jennifer Carney: Yes. So, I said that it's a – these data elements are routinely generated and used for care delivery and available by EHR, so moderate to high.

I was just wondering that preliminary analysis used moderate but I said high. What makes the difference between moderate to high versus high?

Melissa Mariñelarena: That was just due to – usually thought with high on feasibility, if it's a claims based measure, they still require some chart abstraction before it goes into the registry. But again, it's just staff preliminary analysis, so.

Jennifer Carney: OK. Yes.

Shaconna Gorham: All right, additional comments on feasibility?

All right, usability and use.

Jennifer Carney: So, it's not currently used measure that's publicly reported. But it is used in accountability programs, including the PQRS and our ASH maintenance of certification practice assessment module. I think in late 2017, public reporting is planned for CMS.

Shaconna Gorham: OK. Additional comments on usability and use or questions about the measure overall?

All right, well, we have one more measure out. Hand it over to Amber to facilitate.

Amber Sterling: Great. So now, we're going to move on to measure 0378, which is another hematology measure. It's MDS Documentation of Iron Stores in Patients

Receiving Erythropoietin Therapy. Practice in saying that, so hopefully I did it correctly.

Jennifer, Jennifer Carney, unfortunately, you are still our only discussant that has joined the call. So, I apologize for leaning on you so much. However, you are doing a fantastic job. And you're going to be an expert by the time we get to our in-person meeting. So, if you could just give us a summary of this measure. That would be really helpful.

Jennifer Carney: So, this is another maintenance measure with – sponsored by ASH or American Society of Hematology. Looking at those patients diagnosed with low to intermediate risk myelodysplasia who are receiving erythropoietin. And new from the prior endorsement 2012, they actually made some time level duration within the time that iron studies were done 60 days prior to the initiation of the erythropoietin therapy. And the rationale is that those patients that are iron deficient can't respond to erythropoietin, so they shouldn't be receiving it if they aren't being treated for their iron deficiency. That's kind of the gist of it.

So, let's go to evidence unless anyone wants to add something.

Amber Sterling: Does anyone – any other committee members have any questions or comments about this summary before we move on to evidence?

(Dr. Able): This is Dr. (Able). I would just say that this is a very important measure because in older patients with MDS, there often is concomitant reasons for anemia. And if patients just get started on replacement therapy or on a erythropoiesis-stimulating agent without understanding what their iron is they can't make any blood. And we could also be missing some other reason for iron deficiency anemia such as colon cancer or some other lesion on their G.I. tract.

So, it is an important measure. At our place, Dana Farber, we actually have it in the medical record system. That when you go to write for an ESA, it asks you, it shows you what the last iron assessment was. And what often happens in working up for patients with MDS, because they have other measures like

their (MC), the other thing that really look like they have MDS, this gets forgotten. So, it is important.

And also I agree with the change for that it be looked at within 60 days because, often, patients will have a workup for anemia and a long time will go by and nobody thinks to check this again. So, it's both important for diagnostic purposes to make sure you're not missing something else, and for – to the treatments to work because without iron, you can't actually make blood. So, that's my perspective as a clinical MDS person.

Amber Sterling: Great. Thank you, Dr. (Able).

Sam Tierney: Can I just ask one question, Dr. (Able)? Thank you so much. I also agree. This is a very strong and important measure. One, is there a reason to or not, include erythropoietin levels itself as – within that 60 days prior to initiating erythropoietin as a predictor for response?

(Dr. Able): You know, I think that's a reasonable thing to consider as part of the NCCN guidelines. Most of us feel that if the erythropoietin level, the native level is super high, that we're not going to get any response or like to give any ESA. I think that's the issue with that is people often try it anyway. So, it's difficult to know if that would actually guide care.

So, I – well, I very strongly agree with the iron assessment. The assessment of EPO, I think, it's something that we always do and we think is important. But often, we will try an ESA anyway, you know, to hopefully get over their anemia even if it's high. But I think it's – there is definitely evidence for that mentioned for that as well.

Amber Sterling: Great. Thank you so much. So Jennifer, if you are ready, you can move on to evidence.

Jennifer Carney: There are – there have been no new changes to evidence since the last review when it was last reviewed and endorsed in 2012. This remains consensus-based supported by ASCO, ASH and NCCN. The level of evidence is 2A, so it was evidence low.

Systematic review of evidence specific to measure exist. There is quality, quantity and consistency of evidence provided. And the evidence is graded. I did not think that this actually needed to have a voting repeated on the level of evidence.

Amber Sterling: OK, great. Thank you. Anybody have anything to add, any other committee members rather?

Sam Tierney: This is Sam Tierney with the PCPI. I'm sorry. I know you're asking for other committee members. But I wondered if I could just a question?

Amber Sterling: Sure, of course.

Sam Tierney: So, I'm, you know, I'm looking at the algorithm. And I am a little confused just to have this measure and the last measure and even the previous measures that have guidelines supported evidence are getting a low rating according to the criteria. Because even it says, if there no is empirical evidence or there's the empirical evidence but without systematic review and grading of the evidence, then you ask whether or not the evidence that is summarized includes all studies in the body of evidence. And then if, yes, the Steering Committee agreed that the submitted evidence indicates high certainty. The benefits clearly outweigh under desirable effects, and that's the thing. So, I'm just confused as to how these measures are getting a low rating if they're based on guidelines that include empirical evidence and also the evidence is graded.

Female: That was just based on level of evidence was lower level evidence? NCCN describes Category 2A as lower level evidence. That's all that means.

Sam Tierney: OK. I mean, I just – I know that there's a staff level of review that takes place. But I guess I would just, you know, I know that it is framed up to the committee based on their understanding of the evidence. And I know that the physician who just spoke said there is empirical evidence and it is strong to indicate that this is an important measure. So, I just want to sort of the committee to keep that in mind as they consider the rankings of the evidence based from their own review.

Female: Right.

Amber Sterling: Great. Thank you for that input. So, we can go ahead and move on to reliability.

Jennifer Carney: Do you want me to do the performance gap or did I kind of do it?

Amber Sterling: Yes, yes. You can go ahead and do it, sorry.

Jennifer Carney: So the performance, the registry performance rate, there was a wide range and mean of about 54.58 percent in the PQRS average performance rates. The additional study published in 2013 show the lack of concordance of the NCCN guidelines use and community practice by longitudinal assessment within all MDS groups, and that there is no data available on disparities in this area. So, I just want to say there is a performance gap and why this measure is needed.

Amber Sterling: OK, great. Any other committee members have anything to add?

Shelley Fuld Nasso: This is Shelley. This is the one that I got confused in which measure I was talking about, where there was this dip in 2013 that I didn't really understand. And then I also – I don't really understand that how the registry performance mean is still low, but then the PQRS average performance rates are much higher. Maybe I'm not understanding the deference between those two. Can somebody help me with that?

Amber Sterling: Sam, can you respond to that, because I think this is the same for all the PC – or the PQRS measures?

Sam Tierney: Yes. So, I would have say that that is as a result of the PQRS program, so I know that the performance had sort of a dip since it began. And I would say that I don't know how familiar you are with the PQRS program. But it remains a voluntary reporting program. It did in its early years, allow for an incentive for participation, but that is now – it's now in a penalty phase for people who don't participate. So, we expect the rates of eligible professionals who participate in the program to increase. And it in fact has increased over the years, but it is still relatively low.

So, I would say, part of the reason you may see a dip is because more than likely, there are more providers participating in the program now than there have been over the years. And so, the numbers of eligible providers reporting has gone up. And so, the performance rate has gone down a little bit as a result of that.

And the registry performance rate, as to why that differs, so significantly, it is just a year later. Let me ask my colleagues to speak to that in our testing area.

(Deirdre): Yes. So, this is (Deirdre) with the PCPI. So, the registry performance rate section is a different data set than the PQRS average performance rate. So, PQRS average performance rate is data that's pulled from the 2013 PQRS experience report. The registry performance rate comes from a data set that we actually requested and received from CMS for the year of 2014, and it's the same data set that was analyzed for reliability.

And when we analyzed the performance rates from the data set for 2014 for reliability as a part of our signal-to-noise ratio analysis methodology, we have a minimum of 10 events per physician that are included in that sample. And so, some of the physicians might have been kicked out of this sample for the registry performance rate, which may have made it – made the performance rates go down a bit.

Sam Tierney: Yes, and this is Sam One other thing I'll add is that the experience report, I think, the performance rates we get are from performance and claims and registry or any sort of version implementation of the measure. And as (Deidre) described, the data for 2014 is just from the registry implementation of the measure.

Amber Sterling: Great. Thank you so much. Jennifer, does that help answer some of your questions?

Shelley Fuld Nasso: That was Shelley Fuld Nasso. And it does, but it makes me wonder which is closer to accurate? So, PQRS is over stated because only high performers are reporting, and then if the registry is, you know, is it's a more limited subset or if some are being kicked out. I guess I'm not clear like what does this

actually mean as far as the performance then. But it does helped me understand the difference, so thank you.

Amber Sterling: And I think that's a really important conversation for us to have when we discuss this measure at the in-person meeting. That's a really valid concern.

Shelley Fuld Nasso: OK. Well, good. That makes me feel better that I'm not just confused.

Amber Sterling: No, you're not just confused.

Shelley Fuld Nasso: I am confused, but I was confused for a reason.

Amber Sterling: But not just confused.

Shelley Fuld Nasso: Yes. Thank you.

Amber Sterling: Yes. OK, great. This is …

(Crosstalk)

(Dr. Able): This is …

Amber Sterling: I'm sorry. Go ahead.

(Dr. Able): One quick question, just a comment. This is Dr. (Able). So, one thing that may be helpful when you guys have the in-person meeting is we did an analysis using claims data and MDS diagnosis from SEER, in SEER-Medicare that looks specifically add to the measure we talked about before, and this measure to look at performance since 2006.

And we found that only 56 percent of patients had evidence in claims of having had iron assessed before starting an ESA within the same window that you guys are looking for. So, it's just another data source that shows that there is underperformance, likely underperformance on this measure at the national level.

SEER data is not perfect. So, your Medicare data certainly isn't perfect. But since MDS tends to affect older patients, and so it's a good database to look

for these measures, and that's an article that's in (British) Center of Hematology that came out in February online. And it's just – it shows that you can measure these things through claims and that there does seem to be a gap. Although improving over the years, it's still not great by the end of the study, so.

Amber Sterling: Great. Thank you so much. Jennifer, if you are ready, we can go ahead and move on to reliability.

Jennifer Carney: So, interrater reliability exists from a prior review. This study – this measure uses the signal-to-noise ratio which was high. Similar to the last measure, a small sample size because of the more rare nature of this disease. The specifications are appropriate. This measure also used – wait, hold off on validity.

Shaconna Gorham: Yes, you can definitely address the reliability testing though.

Jennifer Carney: Yes, which was high.

Shaconna Gorham: OK.

Amber Sterling: Great. Are there any other committee members who would like to weigh in on reliability or reliability testing, excuse me?

OK. If not, we can go ahead and jump right in to validity.

Jennifer Carney: So, the validity testing method used the phase validity testing. And that based on ASH expert of panel, 89 percent of the respondents agreed or strongly agreed that this would differentiate between good and bad quality.

I just made a note that there is, again, a low number of sample size. But overall, due to the low incidence of this disease, I felt that this made this a reasonable.

And one question I had about, I guess, this issue comes up again about threat to validity. But, you know, make note in the comments from the preliminary analysis that the developer reported that there was a high number of

exceptions. I think 97 exceptions reported amongst the 28 physicians, and that the average number of exceptions per physician was 3.5 with an overall exception rate of 15.8 percent. I guess my question is would this be a threat to validity by having such a high number of exceptions?

Shaconna Gorham: Would the developers like to respond?

Amber Sterling: Or does anyone else from the committee have any thoughts?

Again, I think that's a really valid question and it's something that would be worthy of discussing at our in-person meeting. I think it certainly, as I was re-reading this measure today, it certainly struck me as quite a lot of exceptions. So, it think that that, you know, your question is right on point and it would be something to really get our committee to think about and to answer during our in-person meeting.

Does the developer have any response to maybe why the exceptions were so high or any potential reasons about why?

Sam Tierney: So, this is Sam Tierney. I mean, we all – there is – in this measure, there's only one allowable type of exception. It is system reason for not documenting irons towards prior to initiating EPO therapy. And typically, we might include that because this might have been performed by another physician and the information may not be available, for example, if the patient was under your care and was already receiving EPO therapy. So, there are may be reasons within that. I don't know, (Dr. Able), if you have any perspective as a practicing hematologist as to why there might be people who don't perform this and would have valid reasons why.

(Dr. Able): Yes. No, and I was thinking through because some of those silences, it looked like I was to respond. But I don't really know. I guess the measure is about documentation of iron stores. I think the exceptions would be to documentation, not to seeing or getting iron stores, because this is a blood test.

You know, in our studies, we definitely gave people credit if they got iron stores on the bone marrow, but you can assess iron from peripheral blood. So, I guess, without knowing what the exceptions were, those 27 exceptions, it's

hard for me to know how to address them. But I would think that they might be more documentation because, you know, maybe you received a patient from somewhere else and you haven't done the test but you know it's there. You know, you can see it in their records. I'm just not really sure, you know, what those exceptions would be.

(Deirdre): And this (Deirdre) from the PCPI. Unfortunately, the data set that we received from CMS for the 2014 PQRS data doesn't stratify the reasons or the exceptions, that we just get the number of exceptions that were reported, so.

Amber Sterling: OK, great. Thank you so much. That's helpful in having our committee understand why those exceptions might be so high.

So, we'll go ahead and move on to feasibility.

Jennifer Carney: They've got elements are routinely generated by either the chart EHR and in lab results, so it's quite feasible.

Amber Sterling: OK. Any committee members have anything to add about the feasibility?

If not, we'll go ahead and close this measure out with usability and use.

Jennifer Carney: The current measure is not publicly reported, but currently is used in multiple accountability programs, including PQRS, ASH, maintenance of certification. And similarly, CMS plans to make this available for public reporting in late 2017.

Amber Sterling: Great. Thank you so much. Do any committee members have anything to add overall about this measure or any questions before we move on to our last measure of this call?

Shelley Fuld Nasso: This is Shelley. I had a question. Actually, I have two questions. How does the input and discussion from this call get used or feed into the in-person meeting?

Shaconna Gorham: The developers are on the line, and so they can take the inputs from the committee members and definitely submit additional information for your

questions or points of clarification. We will incorporate that. The developers, again, will be at the in-person meeting. They have approximately three minutes or so to introduce their measure. And at that time, they can again address some of the conversation from the workgroup call. It also gives the committee members the opportunity to kind of hear each other's thoughts and then responses from the developers before the in-person meeting.

Did that answer your question, Shelley?

Shelley Fuld Nasso: Yes, that's great. And then the last question is are we not – do we not have two more measures? Are we not doing the E.D. visits as well?

Shaconna Gorham: We have the 2936 Admissions and Emergency Visits for Patients. That is our last measure for the day.

Shelley Fuld Nasso: OK. Did we do 377 and 378 together?

Shaconna Gorham: We did 37 – 0377 first. 0378, with the one that we just did.

Shelley Fuld Nasso: Oh, OK. I'm sorry. I got confused. Sorry about that.

Shaconna Gorham: It's OK.

Amber Sterling: That's OK. All right, we'll turn it over to Melissa.

Melissa Mariñelarena: OK. So, the last measure is 2936, Admissions and Emergency Visits for Patients Receiving Outpatient Chemotherapy. This measure is developed by Mathematica for CMS. Our lead discussants are David Cella and Jennifer Malin. I don't believe David Cella is on the call. But Jennifer, you are on, correct?

Jennifer Malin: I am. Can you hear me?

Melissa Mariñelarena: We can. Just to say this is a new measure. It's an outcome measure. And I will let you start the discussion, Jennifer.

Jennifer Malin: OK. Thank you. I'm the third of the three Jennifers today. So, this measure I think is an important outcome measure for patients who are receiving

chemotherapy. You know, the rationale basically is that patients getting chemotherapy are frequently hospitalized for side effects. And that both the side effects as well as having to spend time in a hospital or emergency room has a negative impact on patient's quality of life. And that there are things that physicians and healthcare teams and the health care system can do to decrease the incidents of these admissions and E.D. visits.

So, some of those include prescribing appropriate therapy, whether it's antiemetic or white blood cell growth factors, and addressing symptoms upfront, and prescribing the appropriate therapy to medicate them. And then other things include, you know, having systems in place so that patients can get in for timely outpatient care, so that they decrease the incidents of these admissions and E.R. visits.

Actually, I think there has been fairly limited evidence on the ladder. It's been more kind of anecdotal. But there actually was just the study published within the past couple of weeks by the Cleveland clinic that, you know, they implemented a care coordination service and post discharge follow-up visits, and had about an 18 percent relative reduction in the readmission rate. So I think, you know, there is, you know, reasonable evidence for this measure. And if I guess we can win for that section next. But, that's my intro.

Melissa Mariñelarena: Great. Thank you. And because this is an outcome measure, the evidence criteria is a little bit different. We just have to have the committee agree that there is – that the developer provided at least one processes that can impact the outcome. So, if you want to discuss that and see if you agree or disagree?

Jennifer Malin: Yes. So, I think, so they provided a link with sort of structure process and outcome, and the process of being some of the ones that I just described. And I think they've provided, you know, a compelling argument for the link. And I think it's, you know, generally supported. And so, I would agree.

Melissa Mariñelarena: Great. Thank you. So, we can move on to opportunity for improvement and gap and care.

Jennifer Malin: Sure. So, the performance gaps that the developers provide, basically information from – on a hospital-level variation. So, I should have specified upfront, because there are kind of versions of this kind of measure that have been proposed in a lot of different venues that are practice level.

What's being proposed here, as I understand it, is this would be a hospital/facility measure. So, the unit of analysis would be the hospital. And the attributed patients would be the patients who received chemotherapy at that hospital in the outpatient setting.

And so, they present data that risk standard admission rate that the median is 10 percent. Actually the 25th to 75th percentile is actually surprisingly narrow. It's (inaudible) to 10.8 percent, but overall, the range of 10 to 25 percent. The E.D. visit is much narrower. The E.D. visits are – the median is 4.1 percent. The range from the 25th, the inter quartile range is 4 to 4.4 percent. And the overall range is 2.1 to 7.5 percent.

So, I think specifically for the inpatient admission rate, they do provide pretty compelling data on a performance gap. I think the data around the E.D. visit is probably a little more challenging because the, you know, kind of any really sick person with cancer gets admitted. And so, when you're here, it may count. If someone has both in the inpatient admission and an E.D. visit, they count in the inpatient admission. So, the E.D. visits really reflect just those people who went to (inaudible) department and then were discharged home.

Melissa Mariñelarena: OK. Does anybody have any comments, questions?

OK, we can move on do – do you want to talk about the specifications?

Jennifer Malin: Sure. So, as I mentioned, so this is a facility level measure. The numerator is, well, let's say, we do the denominator first. I kind of like doing that. So, the denominator is basically patients 18 years or older with a diagnosis of cancer, who received at least one outpatient chemotherapy treatment at the reporting hospital during the performance period. The exclusions include patients with leukemia, patients were not enrolled in Part A or Part B in the year prior, and

patients who do not have at least one outpatient chemo treatment followed by continuous enrollment in Part A and B for at least 30 days.

So, I think that's pretty straightforward. I guess one, maybe I'll save questions for a minute. But I do have a question for the developer which is increasingly, chemotherapy is also available in oral forms. And so, it seems like this is only – the measure specification only looks at I.V. chemotherapy. So, a clarification around that would be helpful.

Maybe before we get to that question, just the numerator is not all admissions or E.R. visits, that ones that are for – a specific set of qualifying diagnoses, anemia, dehydration, diarrhea, emesis, fever, nausea, neutropenia, pain, pneumonia or sepsis. So essentially, those are diagnoses that are common symptoms of chemotherapy, as well as common symptoms that can arise from cancer.

The timeframe includes 30 days following the day to reach chemotherapy visit, which I think is a reasonable timeframe. Most of these would tend to happen in the first two to three weeks following chemotherapy. So, that's a reasonable timeframe.

And then lastly, you need to be one of the first of second diagnoses on the claims for the admission or the E.D. visits. And if it's a secondary diagnosis, it's accompanied by a principal diagnosis of cancer.

So, I think the specifications I think are all, you know, very reasonable and there – I do have some questions to the developer about that but we can get into those later, maybe when we talk about validity.

Melissa Mariñelarena: OK.

(Crosstalk)

Melissa Mariñelarena: OK.

Christine Holland: This is Holland at Mathematica. Do you want me to address the oral chemo question?

Jennifer Malin: Yes, that would be great. Thank you.

Christine Holland: OK, sure. So, oral chemotherapy is excluded from this measure. So, we don't include it as a chemo encounter or treatment. The primary reason for doing is that it's hard to capture in Medicare claims because you would need the pharmacy data or Part D data which is incomplete and gets very complicated very quickly. So, that was one driving factor. And then discussing it with our expert panels, they said that adverse events following oral chemotherapy are rare and don't need to be a focus of this measure.

Jennifer Malin: I would disagree strongly with that, honestly. So, I mean just I think the measure – I mean I think I hear the issue you're talking about. But, you know, certainly for, you know, future iterations, it would be good to come up with a way of including the orals. They're increasingly common and they're probably going to end. You know, if they aren't today, they're going to easily be 20 to 30 percent of all chemotherapy going forward, and they have just as many side effects.

In many cases, it's actually more challenging to manage with side effects, because with the I.V., at least you know people are coming in every two to three weeks. And orals, people can often get left off on their own.

Christine Holland: OK, that's good to know. And …

Jennifer Malin: And then, you know, the treatment related mortality for some of the new oral therapies are in the, you know, in the high single digits. So, it's clearly an issue.

Christine Holland: Yes, that's a good idea for expansion in the future.

Melissa Mariñelarena: OK, and are there any other questions?

Danielle Ziernicki: I just have a question, this is Danielle Ziernicki, regarding the exclusion criteria and potential expansion beyond patients with leukemia to other diseases that are also can cause, you know, the underlying disease could be potentiated by chemotherapy, but also you may see some of the primary

diagnoses such as dehydration, et cetera, outlined here. Have you considered expanding the exclusion criteria?

Jennifer Malin: So, I'll let the measure developers. I think I just want to clarify. So, leukemia is excluded for the measure. So, it would include, as I understand, all the other cancers except people with leukemia.

Danielle Ziernicki: That was my understanding too, Jennifer.

Christine Holland: This is Christine Holland. I thought, again, yes, that's correct. So, you have to have a diagnosis of cancer during the period to be in the denominator, and then we exclude patients with leukemia.

Jennifer Malin: And is that any leukemia or only acute leukemia?

Christine Holland: Any leukemia.

(Melissa Mariñelarena): And have you considered expanding – I guess this question is to Mathematica. Have you considered expanding that exclusion criteria beyond leukemia?

Christine Holland: So, I'm assuming you're asking about other cancer types that might want to be excluded during development. We also include – did some analyses, looking at whether to exclude lymphomas. But ultimately, decided not to, that they had similar rates to the other cancer, similar treatments. And if they were getting their treatment at an outpatient setting, their care should be managed similarly.

So in that regard, we considered one other, but aside from that, no. The patient does have to have cancer to be in the measures, so other conditions that would have someone getting chemotherapy aren't considered here.

Jennifer Malin: I guess just one other question. It doesn't appear that you include any other infections besides pneumonia and sepsis. Is that the case?

Christine Holland: Yes, for the numerator outcome.

Jennifer Malin: And that is there a reason for that?

Christine Holland: So, I can speak to the historical perspective of how we got here, and then maybe I'd invite (Joe) who – or Dr. (Ross) who has been consulting on the development to see if there's additional clinical rationale for it.

But the reason that we landed here is that historically, when we were developing the measure, we were looking to capture patients with neutropenic fever. It turns out, there's not an ICD-9 code for neutropenic fever. So then, we split that into neutropenia and fever. And then neutropenia is actually often not coded on the claims, particularly as the primary reason for diagnosis because it requires lab results, and usually the inspection is what gets coded. So then, we expanded it to include sepsis and pneumonia being the most prevalently related infections.

(Joe Ross): And this is (Joe). And I just (inaudible). In conversations with, you know, various oncologists and the primary concern around this sort of preventable management or preventable admissions and E.D. visits related to neutropenic fever. And so that's why there was the focus on that and not just on any infections, so any admission …

(Crosstalk)

Jennifer Malin: Yes, I hear you. I think the – and, you know, this is kind of your worst nightmare which is the committee member presenter, someone who's published widely on using claims data for identifying fever and – febrile neutropenia in claims. So, this is probably a little – I'm too in the weeds potentially.

But the challenge is just basically febrile neutropenia or fever and neutropenia tends to be kind of it's the – it's a diagnosis of exclusion. So, someone who's getting chemotherapy presents to your E.R. with a fever and a low white count. And so you look for a urinary tract infection, urosepsis, pneumonia. And then if, you know, you do blood cultures. And if you don't actually find a specific diagnosis, then you're left with fever and neutropenia. But if you find

a more specific infection related diagnosis, that's usually your first or second diagnosis.

(Joe Ross): And just out of curiosity, as we think about how to improve measure going forward, beyond the pneumonia and sepsis. So, I guess, are you suggesting that we would not look for pneumonia and sepsis or …

(Crosstalk)

Jennifer Malin: No, I think you would include those. But you'd probably want to include things like bacteremia, line infections, you know, pyelonepthritis, you know, other infections that results from, you know, that are basically occur during someone, you know, when they're neutropenic.

(Crosstalk)

Jennifer Malin: I mean I don't think it limits that validity of the current measure, but it will also probably helps you out with the sample size, which is sort of one of your issues later on.

(Joe Ross): I agree. I'll also just note back even though Christine addressed this, which was the issue of oral chemotherapy. And it's exactly the issue you raised, which is the challenges of attribution for prescribing oral based therapy as opposed to the infusion based. And so, the first focus was on hospital outpatient departments that are providing infusion based chemotherapy.

Jennifer Malin: Thank you.

Melissa Mariñelarena: Are there any other comments, questions? If not, we can move on to reliability testing.

Jennifer Malin: OK. So reliability testing, the developers used at the intra-class correlation signal-to-noise method, and recommend a minimum number of cases to achieve reliability of points for our greater – this part, I may need some help with because I was a little confused because the methodology says that it was 942 hospitals with at least 60 patients. And then down below, they end up

saying a minimum of 25 patients, so that was a little confusing to me since it seemed like they had excluded that.

But basically, the inpatient admission reliability was 0.4. The E.D. visits appeared to be below their threshold at 0.27. So, you know, perhaps – and then the reliability algorithm, I think reflecting that showed that the inpatient admissions were moderate and the E.D. visits were low. So, it might be helpful for the test developers to kind of comment on that because I probably confuse things here.

Christine Holland: Sure. So, this is Christine Holland again and I will start. But then, I'll invite my colleagues and statisticians on the line to feel free to chime in if I misrepresent. But we did do two different types of reliability testing. The first one we did was a test-restest. And we took the one year of data that we had and we split it in half to compare half of the performance to half of that performance, which actually ends up giving us only like a half of it, equivalent to what would be a half of year of data, which is one of the reasons the reliability may be underestimated here.

And then implementation, we would have a full year of data, which may increase the feasibility. In addition, we did that second level of analysis to look at the minimum case counts. And using a full year of data landed on a minimum case count of 25. And I know, (Fe), who is on the line can explain a little bit of a nuances between the different counts and the approaches that we did.

(Fe): Sure. This is (Fe). So Jennifer, regarding to your question of the minimum case count using the signal-to-noise method, as well the reliability testing using the test or restest approach. So, here is the nuances there. So when we do that reliability testing, and we – first of all, we tried different approaches. And the reason we finally launched the test and retest approach is because, you know, over scenario. So for instance, we use this measure. It's a risk adjusted.

So, they are to estimate some reliabilities. So slightly from the framework, it's different from, for instance, other process measures where they have the

(banner) outcomes. And, for instance, could use the beta-binomial method to estimate the signal-to-noise reliability. So, that's why the main reason we used the test-retest approach here.

And the reason, during the test, we choose the minimum case comp in the calculation using 60 is to reduce the small sample size influence to the reliability estimate. And the other, to have at least a 30 case based within each the test or retest, that's why we chose the 60 in the testing.

And the calculation of the threshold is really the rationale is to inform the policymakers. So for instance, you've used this measure in certain programs and what other caveat, the nuances, we need to incorporate in terms of the small sample size. And we used the signal-to-noise approach here. And basically, that's based on formula could estimate the minimum case count. And we get, for instance, if we set the threshold as 0.4, we get the minimum case count. I think it's 25 for E.D. measure. And the 25 and the 30, they may look, first, a low case. It's not consistent. But the nuance is that in the testing, what we based in the reliability estimate, what we used is all the hospitals, there are case count is above 60.

So really, in this calculation, as we mentioned in the test, the document, there are around 900 hospitals out of this over 3,000 hospitals in the reliability estimate, whereas when we considered minimum thresholds for the case count estimate using the signal-to-noise framework, we include all the hospitals there.

So in that regard, so for instance if we have all the 3,000 hospitals, their signal variation could be larger. Therefore, in that regard, over estimated, the minimum case count is work for over 3,000 hospitals could be lower than, for instance, the 30 or equivalent to a 60 test and retest, the in total in our estimate. So, that's the explanation for some of the disparity of the results. That overall, this – I think the methods should align.

Jennifer Malin: OK. So, let me just see if I understand. So, your reliability results for E.D. visits is 0.27. But you say that it's a minimum of 20 patients. So, am I to understand that you've got a point to – that with the minimum of 20 patients,

that the E.D. visit rate would be – would have an ICC of 0.4. So, am I to understand that the reason that the score was 0.27 is that a number of hospitals included in your estimates did not have at least 20 patients then, and that's why our score was lower than 0.4?

(Fe): Yes.

Jennifer Malin: OK, so …

(Crosstalk)

Jennifer Malin: So, my follow-up question I guess is out of the 3,765 hospitals, how many would meet the minimum criteria for hospitalizations, and how many would need the minimum criteria for the E.D. visit rate?

(Fe): I think we did that analysis, yes, (over) testing. But I need to double check whether we included this information in the testing form.

Jennifer Malin: OK. Yes, I think that would be really …

(Crosstalk)

Jennifer Malin: … important for the committee to have because if your setting – you know, if in your subset that you used for this analysis where you, you know, you – it was just the hospitals that at least had 60 patients, you would still have to subset it even more in order to get the minimum ICC for E.D. visits to be 0.4, then it seems like you're getting to a pretty small set of hospitals that the measure would apply to.

(Fe): Yes.

Melissa Mariñelarena: Jennifer, this is Melissa. Do you – are you OK with a threshold for reliability of 0.4?

Jennifer Malin: I don't really know. I guess I don't really have enough context in terms of what other measures typically have.

Melissa Mariñelarena: I mean, we usually – I guess it depends. We usually say 0.7. We've also heard for blood sample methodology, that 0.4 is sufficient. I mean based on the – what it says here, like the 0.41 still falls under weak. So Christine, I don't know if like you can offer some insight as to why 0.4 was chosen.

Christine Holland: Sure. I can try and then I'll offer my colleagues to chime in.

Melissa Mariñelarena: Thanks.

Christine Holland: So, a few things, the guides from (events) that you cite in your worksheets does classify it as weak. When we did our testing, we used (Cohen), I believe, is the classification which is if it's under three, then it's weak. Between three and five, it's moderate. And above five is strong reliability. Which is why in our quorum, we've classified our results as moderate because it fell in that – for that classification for that. I think it aligns with the ICC reliability testing. So, we did use a different guide of somewhat you use.

And then furthermore to the question of context or what other measures have, in our submission form, we did include some references to other NQF-endorsed claims based outcome measures. That – and they're ICC reliability scores which range from like 0.33 to 0.38. So, the results that we're finding here do align with several other NQF-endorsed outcome claims based measures. And then we chose 0.4 for the minimum case count thresholds to align with CMS policies for their analysis for minimum case count and implementation considerations.

Melissa Mariñelarena: Thank you. And can you send me your (Cohen) of what do you – the one that you use and I will include it in the measure form for the committee.

Christine Holland: Of course.

Melissa Mariñelarena: Thank you. Any other questions? If not, Jennifer, we can move on.

Jennifer Malin: OK, so moving on to validity. So, the method used was face validity by external groups. And, you know, I think the – you know, we've discussed some of these issues already. But, you know, there is I think kind of both within their process described here as well as, you know, the fact that versions

of this measure have been proposed by a number of different provider organizations and some patient organizations. So, I think that support the idea that there's general face validity for this concept.

I think the kind of, you know, threats to validity, we've touched on already with the oral chemotherapy. I think less so in terms of validity but more and that it would help potentially with the reliability is that there may be opportunities to capture some more diagnoses. But the main potential threat to validity I think is the oral chemotherapy. Otherwise, I think we know that the developers have done a very nice job of summarizing the validity considerations.

Melissa Mariñelarena: Great, thank you. Any comments, questions?

OK. We can move on to the statistical model.

Jennifer Malin: So basically, it might be helpful for the test developers to, you know, describe the model. It was pretty complex and I'm not sure I could summarize it easily.

Melissa Mariñelarena: OK. Christine, do you want to take that or pass it off to one of your colleagues?

Christine Holland: So, my usual of trying to start myself, but then invite my colleagues to chime in if I go astray.

Melissa Mariñelarena: OK.

Christine Holland: So for our risk adjustment methodology. So this measure have two outcomes as we showed – as was shown earlier in the performance table. We look separately for whether a qualifying inpatient ad mission happened. And then, if they aren't admitted, we look for a standalone E.D. visit, a qualifying stand alone E.D. visit.

So because we report two rates, we also developed two risk adjustment models using the same starting point of possible risk factors, and then running a typical regression backward selection with the threshold of 0.05 to select the variables.

So ultimately, we ended up with some demographic variables, including age and gender in the models. As well as cancer type, we've grouped those into nine different cancer type categories. Working with our expert panels for clinical similarities on the cancer types to group them, as well as I think nine different potential co-morbidities. Again, working with our expert panels to kind of group this up into larger groups, as such, like cardiovascular disease rather than specific diseases within that. So, there are kind of nine larger groupings.

And then ran our backward selection and end up with two models. One for the inpatient admission and one for the E.D. visit with similar characteristics. One of them ending up with 20 variables, and one ending up 15, a couple of the cancer types and co-morbidities were significant in one model and non significant in the other.

I don't know how far to keep going.

Jennifer Malin: No, that was very helpful. I guess is – and was the list of the variables included in the models provided, or just basically the outcome rates at – in Table 2?

Christine Holland: In the testing form, should be both the list of all variables considered, as well as the list of all final variables.

Jennifer Malin: OK. And then the other part I was a little confused by is the observed admission rates, the maximum as 100 percent?

Christine Holland: Yes. So this, the analysis again weren't limited by sample size. So you could – I can't remember the specific to the top, a little bit theoretically, that could be a hospital that had one patient that they administered chemotherapy …

(Crosstalk)

Jennifer Malin: So there was no minimum number of patients then in these analyses base?

Christine Holland: Correct.

Jennifer Malin: Because your – because up above, when you did the – but when you did the overall rates of admission, it was like 25 percent in table, like, the range in the performance gaps which looks like it was all the hospitals, the inpatient admission rate range from 6 percent to 24.9 percent. So, I was confused to have down here was 100 percent.

Christine Holland: OK. I can go back and look at the forms. And if we did it, consistently implement or not implement minimum case counts, that we can make sure to do that in an updated format. Off the top of my head, I would've thought we did it, but I will have to go back and check.

Jennifer Malin: OK, and I don't – I mean, I don't – you know, I guess I'm just trying to prove that I actually read through this in detail. I don't know that it necessarily matters, but it would just be helpful to have the inconsistently – inconsistency explained.

Christine Holland: Right, it's helpful. And that means that you know this form better than I do right now.

Melissa Mariñelarena: And we'll make sure it wasn't something that I messed up too.

Christine Holland: OK.

Melissa Mariñelarena: Because could've been it.

So we want to talk briefly, we have about six minutes and we still need to include to have like comments. If we want to talk briefly about SDS factors that were considered in the risk adjustment model. Jennifer, Christine, (me).

Christine Holland: I'm sorry. Did you want me to expand more on the factors that are in the risk adjustment model?

Melissa Mariñelarena: We could just talk about what we did for SDS. This is an outcome measure that it was required. So, if you want to talk about briefly the SDS factors that were considered outside of what was included in the model.

Christine Holland: OK. Also, real quick, while you were talking, this is why I wasn't listening, I was flipping through the form and I think one of the difference, I'll confirm that's in writing, right? But I think one of the differences in the tables that we are looking at is that one was from risk adjusted results, which the range would be different from the observed risk rates later presented. But anyway, I'll go back and check all of that for you guys.

OK, and then to the SDS. So yes, under the NQF trial period, we did expand our analysis to consider some sociodemographic risk adjustment factor. Specifically, we included factors that were available and measure claims data, as well as an American or a community survey index encounter, like in neighborhood index encounter at the ZIP code level that following an (AHRQ) methodology of combining that for another factor.

So we looked at race and a risk factor – sorry. And then when regarding the methodology, what we did was after we had finalized our model which we just talked about. What we did was we looked at the patient level to see across the patients, is there a difference in care for these certain factors?

And then at the hospital-level, is there a variation in the hospitals treating these different types of patients and their performance on the measure? We did find that. And then thirdly, we looked – when we add them to the risk model, is there a difference in how the model fits or how the hospitals are ranked, which we include or don't include these?

So at the patient-level, we did find that there is a difference across these risk factors. However, so like dual eligible, which we use as a proxy for income. Oh, I think I forget to mention that one. But dual eligible which we use as a proxy for income were more likely to be admitted than non dual eligible. Black patients were more likely to be admitted than non-black patients. And lower SDS indicator patients were more likely to be admitted than higher index.

However, when we moved it to the hospital-level, there was no between hospital effects that could be seen. So for example if you looked at hospitals that were treating mostly black patients, their performance rates were similar

to hospitals that treated mostly non-black. So at the hospital-level, there wasn't a difference in performance rate based on the patient mix.

And then when we added it to the risk model, we also found that there was no variation. There was – the fit wasn't any better with or without the factor. And the ranking of hospitals was similar with and with and without these factors. So, based on these findings that we decided including SDS factors were not necessary for our model.

Amber Sterling: Great, thank you. Jennifer, do you have any questions or anything that you'd like to add?

Jennifer Malin: No, thank you.

Amber Sterling: OK. We can move on since we just have a couple of minutes, quickly.

Usability and use, feasibility, so we can move on to – we have one minute for public comment.

Operator: At this time, of you'd like to make a comment, please press star then the number one on your telephone keypad.

And there are no public comments at this time.

Melissa Mariñelarena: Great. Jennifer, would you like to add any last thoughts while we have a minute left?

Jennifer Malin: No, I just – I'd like to thank the folks at Mathematica. I thought they did a very thorough job. So, thank you.

Melissa Mariñelarena: Great. Thank you. Does anybody else have any questions, comments? OK, we have.

Shaconna Gorham: For the Standing Committee members, we just want to ensure that you all have registered and received all of your travel information. Our in-person meeting is weeks to come so around the corner. If you are having any difficulties registering or booking your flight, please let us know so that we can assist you.

For the developers on the call, thank you for attending this call. And if you have information, additional information you would like to submit to us, please do so through the cancer project mailbox.

For our lead discussants, we thank you. You did an excellent job today. If there are no other questions, I just want to take one minute for any questions for Standing Committee members.

OK, hearing none, we will end the call today. Thank you for your participation.

Melissa Mariñelarena: Thank you.

Shaconna Gorham: Have a good day.

Female: Thank you.

Female: Thank you.

Female: Thank you.

Operator: Ladies and gentlemen, this does conclude today's conference call. You may now disconnect.

END