National Quality Forum
Moderator: Sheila Crawford
August 27, 2012
1:00 p.m. ET

Suzanne C. Theberge: Hi, everybody. And welcome to the Neurology Endorsement Maintenance Phase I Post-Comment Call. This is Suzanne Theberge with NQF.

And I'd like to start the call by running down a roster of our Steering Committee Members and our Developers, so we know who's on the line. And then, before I do that, just a couple of housekeeping items – if you are not speaking, please put your phone on mute, so we don't have background noise and please don't put us on hold during the call as we'll hear your hold music.

OK. So, and also the webinar portion of the call – if you're dialed into that, you can chat us a question if you want to bring up a point and you are having trouble getting us, a word in edgewise, just send us a chat message and we'll make sure that you are called on.

OK. With that, let me go down to Steering Committee Roster. David Knowlton?

David Knowlton: Here.

Suzanne C. Theberge: Great. David Tirschwell?

David Tirschwell: Here.

Suzanne C. Theberge: Anna Barrett? OK. William Barsan?

William Barsan: Here.

Suzanne C. Theberge: Jocelyn Bautista? Ramon Bautista? Gwen Buhr?

Gwen Buhr: Here.

Suzanne C. Theberge: Gail Cooney?

Gail Cooney: Here.

Suzanne C. Theberge: Tina Cronin? Jordan Eisenstock? Risha Gidwani? David Hackney?

David Hackney: Here.

Suzanne C. Theberge: Let's see. Greg Kapinos? Michael Kaplitt? Daniel Labovitz? Terry Richmond?

Terry Richmond: Here.

Suzanne C. Theberge: Jack Scariano? Raj Sheth? Jolynn Suko?

Jolynn Suko: Here.

Suzanne C. Theberge: Jane Sullivan?

Jane Sullivan: Here.

Suzanne C. Theberge: Fred Tolin?

Fred Tolin: Here.

Suzanne C. Theberge: Mary Van de Kamp, and Salina Waddy?

All right. Is there any Steering Committee Members on who didn't introduce themselves? OK. I know a couple of folks said they might be dialing in late due to clinical duties, and I know a lot of people are on vacation this week.

So we appreciate those of you who are here joining us. And now, I'm just going to run through our measure developers. AMA-PCPI are you on the line?

Diedra Joseph: Yes, we are. This is Diedra Joseph and Jennifer Trajkovski.

Suzanne C. Theberge: Great. OK. AHRQ, are you on the line? OK. Esley Schwam, are you on the line?

Female: Yes, we are. I'm Dr. Fonarow and Dr. Schwamm is on the line as well as, I believe, Dr. Smith will be calling in.

Suzanne C. Theberge: OK, great. CMS and Yale, are you on the line?

Susannah Bernheim: Hi, this is Susannah Bernheim from the Yale team. We are here.

Suzanne C. Theberge: Great. Thank you. And Joint Commission, are you on the line?

Ann Watt: Yes, this is Ann Watt.

Suzanne C. Theberge: Great. Thank you.

Well, with that, unless anybody has any questions, I'm going to turn the call over our co-chairs Dave and David.

David Tirschwell: OK. This is David Tirschwell speaking. Thank you all for joining us today. And for the Steering Committee Members, the call's going to be organized based on the agenda which also correlates to – what was the other document called? The memo. Is that right?

Suzanne C. Theberge: Right. The briefing memo that we sent out on Friday.

David Tirschwell: Briefing memo – we'll be running through the major themes that came across in the public comment period, and we'll entertain discussion by all the Steering Committee Members.

In our preview of this call, we wanted to make sure that people knew that we will not be revoting necessarily today. In fact, we'll be discussing whether people want to revote or not on any of the measures, and that – if people decide we want to vote and probably be a low

threshold for voting again, then that would happen over the next week or so via an online voting survey. And the NQF would also help summarize our discussion today for us to have available at the time of such revoting.

And then, we'd also remind the developers and anybody else that is on the call that this is mainly a Steering Committee Conference Call. And we request that you respect that and only speak if a specifically asked a question, and try to constrain your comment to specifically answer that question.

OK. Were there any other housekeeping issues before we get started?

Suzanne C. Theberge: No, I think you have it Dave.

David Knowlton: I think you're right Dave.

David Tirschwell: OK. So then, on the briefing memo, I'm on page two. And it summarizes a number of things including the fact that 53 comments came in from the public and NQF measures.

NQF has coalesced these comments into three major themes that we'll going over, but if anybody wants to bring up things that don't mentioned in these three major themes, committee members should feel free to do so. We've included that in the agenda under additional discussion on comments and responses towards the end.

And so, I'm going to just jump in to the first major theme which is feasibility. And there were 13 specific comments regarding the feasibility of several measures, most of these comments were along the lines that it may require burdensome electronic health record data extraction on medical chart review. And the 12 measures that this comment was made about are listed there on the bottom of page two and the top of page three. Continuing on page three, there was also a repeated concern and thank you.

I guess, if you're following on the website they're scrolling through this as I'm reading along, so you can follow along on that as well. There was the concern that the measures might be difficult to implement from administrative claims, but in fact, there are some CPT II codes which address some of these and the joint commission measures were not specified for administrative claims.

So, the proposed committee response was – yes, go ahead. Somebody have a comment?

The proposed committee response was that while these measures may require a fair amount of data abstraction, the measure met and continue to meet NQF's feasibility criteria. Does anybody object to or want to modify that as a proposed committee response? Can people here me?

David Knowlton: Yes, we should probably (inaudible) David.

Male: Oh, yes.

Male: I think that's fine.

David Tirschwell: Oh, OK, great. Thank you. I thought I might have been alone there for a second. OK, well then, let's – I don't know that – do we need to necessarily vote or otherwise affirm that we want to move ahead with the proposed responses, or that's already being noted and tracked by NQF staff?

Suzanne C. Theberge: We're noting and tracking, so as long as no one has any difficulties with that, we'll just proceed.

David Tirschwell: Fantastic. And then, there was one final comment about feasibility concerning the time to intravenous thrombolytic therapy, and again, bringing up that it would be hard to do from administrative claims alone.

Apparently, there was an error in the original submission and this was never meant to be captured by claims data. And so, the measure itself has been revised to specify that it's from – I guess, chart review or registry data only. OK.

David Knowlton: I think that's fine. That sounds fine too.

David Tirschwell: Anybody have any other comments about feasibility, or does that change anybody's mind about any of the measures that have been proposed. I'm going to take silence as agreement.

David Knowlton: (Inaudible) OK. Yes.

David Tirschwell: OK, all right. Dave K., do you want to run through theme II harmonization or –?

David Knowlton: No, why don't you go ahead David. You've got this going just fine.

David Tirschwell: OK. I'll continue along. And then, in a few minutes, we'll get to the thornier issue about strokes that vary which may take up much of the call, based on the amount of feedback that we got that.

So, theme two was about harmonization. There were three comments suggesting that numerators, and denominators, and exclusions, and timeframes for measure 244 and 441 be harmonized. 244 is that stroke rehabilitation services were ordered, and that's an AMA-PCPI measure. And 441 is that Stroke 10 measure from the joint commission where it's assessed for rehabilitation as opposed to rehabilitation services being ordered.

The differences are outlined in the table at the bottom of three that's scrolling away as well as on the top page four in the memo. There are some differences including most notably that AMA measures as – I think, most of their measures are at the clinical level versus at the facility level.

The committee did not identify any harmonization issues to be developed by – I mean, to be addressed by the developers. Nonetheless, the developers did respond. The AMA-PCPI, suggested that their measure was fine as is.

The joint commission suggested, and they repeated this for some other measures, that they've been working hard to – with the AMA-PCPI and other groups including the Heart and Stroke Association and the CDC with Paul Coverdell Registry to try to harmonize measures. Then

they will continue to do so in the future. And, I guess, we can only hope that there'll be some more success in the future.

An initial proposed committee response for this harmonization issue was that while the committee agrees that the measure should be harmonized, the committee members recognize that the measures are specified for different levels of analysis, clinician versus facility thus may require somewhat different specifications.

And, I guess, I would just add that while there may be some differences in the specifications, it's the actual data point that need to be collected, overlap 80% or so, or one is a subset of another, it would certainly allow for potential cost savings at the hospital and/or clinician level.

So, I certainly would like to endorse continued and perhaps even more aggressive efforts at harmonization between these major developer groups. Does anybody else have comments?

David Knowlton: This is Dave Knowlton. David, were you suggesting a change to the recommended response or are you OK with that?

David Tirschwell: Well, I guess, this is a point of procedure – does the NQF or can the committee make more specific recommendations like that? And I guess, more importantly, do they have any, any binding power?

Helen Burstin: Hi, Dave. It's Helen Burstin. It would be fine for the committee to make specific recommendations and if nothing else, they will be looked at when the measure is next up for maintenance.

David Tirschwell: And is the cycle for maintenance, for measures pre-established or can the committee weigh in on that as well?

Helen Burstin: For all measures in general, it is every three years unless there's a change in evidence or some information out there in the community. We have been having committees fairly, routinely asking for some harmonization efforts within one year by the next annual updates. So, that's certainly within your purview as well.

David Tirschwell: So, I would propose that we add a comment to this harmonization issue. And I don't know if we can add it to other measures that were proposed as overlapping and whether there were harmonization issues in the draft report, suggesting that, I would certainly be interested in getting an update at one year's time and evaluating the progress. So, anybody second that?

David Knowlton: This is Dave Knowlton. I agree with you, David.

David Tirschwell: Any objections out there in the committee?

Female: No.

David Knowlton: Sounds good.

David Tirschwell: OK. Great. Well then, if there are no other comments – and again, I guess, I just note that especially as we get into the stroke severity and risk adjustment theme, if there are many

people that are talking and you haven't had a chance and do want to say something, I think if you just make a quick comment in the chat box, NQF can help us make sure that everybody who has a comment on the committee has their chance to make such a comment.

OK. We are pretty much right on time so far. So, theme three and perhaps the theme that generated the most words in the public comment period was that related to severity of stroke and risk adjustment models. And specifically, there were 18 comments most of which suggest the great value of the NIH Stroke Scale as a measure of stroke severity in a very powerful predictor of outcomes and of course, we're referring to stroke severity at the time of presentation to a hospital.

And then – the comments related to the three risk adjusted outcome measures listed there – 0467, which is the acute stroke inpatient mortality rate, IQI 17 from HARQ. Then there's 2026, which is the hospital 30-day, all-cause, risk-standardized mortality measure referred to as RSMR in acute ischemic stroke only from CMS and Yale.

And 2027, which is sort of a partner measure, but it doesn't look at all-cause mortality, it looks at all-cause risk-standardized readmission rates following again, just acute ischemic stroke from CMS and Yale.

Most of the comments that came out, I think, many of them are related to the mortality measures and quoted the recent article by Fonarow and colleagues working with data from get with the guidelines that had been merged with Medicare data.

And they quote – in this memo, there's a quotation of the conclusion of the article, and I'll just go ahead and read that. The quotation at the end from this paper with that adding stroke severity as assessed with the NIH Stroke Scale Score to hospital 30-day mortality model based on claims data for Medicare beneficiaries with acute ischemic stroke is associated with substantial improvement in model discrimination and changes in mortality performance ranking for a considerable proportion of hospitals.

These findings suggest it may be critical to collect and include stroke severity for optimal hospital risk adjustment of 30-day mortality for Medicare beneficiaries with acute stroke. And I'm sure the authors feel that would like apply to risk adjustment for all patients with ischemic stroke, although they are only able to assess the importance in this paper as it relates to Medicare beneficiaries.

So, there are a number of responses; a response from AHRQ which is quite long; a response from CMS and Yale which is also quite long. And I can go through those, but I guess, I would like to open it up preliminarily to committee members to get early comments on how you feel this article and these other responses that we got back may or may not have changed your mind about these measures.

OK. Honestly, I find it –

Female: I'm sorry –

David Tirschwell: nobody has an opinion –

Female: This is –

David Tirschwell: – go ahead.

Terry Richmond: Yes. This is Terry Richmond. Can you hear me now?

David Tirschwell: Yes.

Terry Richmond: I have a comment. We had a lot of discussion about this in our meeting and the importance of the – adjusting for the stroke mortality. And I was on a little workgroup with these measures, and I read all these responses and I went back to my original notes.

I think it is compelling the importance of risk adjusting for stroke severity as – at the time of (intake) to the hospital. I'm still; however, quite comfortable with in the hospital mortality measure, the 467 I believe it is – because, well they don't us the NIH as a stroke – the stroke scale, they do use clinical markers to adjust for stroke severity.

So, it takes it from a different approach, but they are incorporating it. And I think, from my perspective, I still feel quite comfortable with that. I'm not quite – I don't – I personally don't have quite the same level of confidence with the two 30-day outcome in terms of adjustment. But they're two very different approaches and different measure, but the in-hospital one does, in a different way, certainly adjust for clinical markers of stroke severity.

David Tirschwell: And if I'm recalling some of those documents, they comment on – there's a code for coma, I guess, and hemiplegia which they kind of role out as – although, as you say, not the NIH Stroke Scale, certainly, not – it certainly captures of the stroke severity information as you suggest. Is that right?

Terry Richmond: Correct. Correct. And they gave us quite a bit of detail that's more than in this response memo that – in terms of their – the four different levels of how they really risk stratify. So, I can't really speak for them, but I was – I think, when I take that whole package together I'm much more confident that they're incorporating this sufficient level within the feasibility getting data like this of dealing with stroke severity.

David Tirschwell: OK.

Daniel Labovitz: Daniel Labovitz here. I remain as uncomfortable as ever, in fact, more uncomfortable than I was when we first reviewed this. And I had a – I felt like I wasn't well prepared when we had our meeting for dealing with these measures. It's – I've really only learned as a part of our process how to start to think about them.

And I felt that by having – part of it has to do with the NQF process where basically, it's the sponsor of the measure who makes – or, the pro and con arguments. And I felt like the con arguments were really not well presented. And I found the Fonarow paper to be highly convincing.

We also have the same problem – we have the same problem, really for mortality and for readmission rates. And I disagree that these administrative codes substitute for clinical measures. If I write that the patient has hemiparesis from a stroke using a 438 code, that can be hemiplegia or it can be subtle and nonclinically significant weakness.

There's no way to tell from these administrative codes how bad the stroke is. Yes, a coma might kill you, but it's – that's extremely crude. And I feel like it doesn't make sense to me that somehow we could get away without a better clinical measurement. And I think, bad data is worse than no data.

Gregory Kapinos: I concur – Gregory Kapinos.

William Barsan: Dave, this is Bill Barsan. I think that Daniel said it pretty well. I think most of us would feel a lot more comfortable with the NIHSS as an instrument that pretty well is very good at predicting outcome. And, yes, I think it's – I mean, I think it is an issue.

And we talk about that before, where the other measure that was going to recommend that NIH Stroke Scale be collected as part of the data. You know, it's hard to make a case that just collecting NIH Stroke Scale is actually a quality measure, but it certainly keys into this measure very much.

David Tirschwell: Right. So, I mean, it seems like everybody's in agreement that all of these measures would be better off if the NIH Stroke Scale was available for risk adjustment. I think, even the measure developers would likely agree to that and, in fact, I think, I read that in many of their comments.

The reality is that it's not and that it would likely – I don't know, my guess is that the way these things move, it's unfortunately probably slow. And if that were the case that these measures wouldn't go forward until the NIH Stroke Scale was available, it would mean probably, I would guess, at least another three-year cycle before that was available.

So, I think, as Daniel suggested, if you consider these measures as potentially producing bad data and that's worse than no data at all then you might want to consider voting against endorsement.

If you think that these measures, while they are lacking the NIH Stroke Scale Score, do have some value and that, that it's good enough for now with, of course, the hope that it will be improved in the future perhaps, based on the NIH Stroke Scale, then perhaps you could consider still endorsing. And I guess, the final votes still came out in favor of all the three of these measures at the end of our committee meeting in person back in June.

Other comments from the committee about these measures? I know we started with the AHRQ acute stroke mortality measure, but I think there's somewhat of a fluidity in many of the arguments. The memo that we have – I guess, I'm hoping that all the committee members have had a chance to review it so I don't have to read it out in great detail. But, does anybody want me to read through it?

David Knowlton: No, I don't think you need to. I think we've read it.

David Tirschwell: OK. And I guess, I would open to the committee – would anybody like – because I think, well, certainly, at least, for the CMS and Yale risk adjusted 30-day measures for acute ischemic stroke, we heard that those measure developers are on the line. Does anybody have a specific question they would like to ask to the developers?

Risha Gidwani: This is Risha Gidwani. Are you able to hear me, as having –

David Tirschwell: Yes, yes.

Risha Gidwani: Hi. So, I think when we look at these measures, it would behoove us to separate them out and look at the discriminative ability of each one of the measures individually. When we're looking at the AHRQ in patient mortality, from the documents that the developers originally sent to us that we reviewed in our in-person meeting, their c- statistic is 0.894 which means that they have almost 90% ability to distinguish patients who died than who didn't die. And they're able to do this without the NIHSS information.

When we look at the CMS models, they're discriminative ability between either patients who died, did not die, or patients who were readmitted, were not readmitted is much lower. And so, when we talk about the inclusion of this new clinically related variable, I agree with the comment that were made that it provides a level of information that goes beyond with the administrative data can give us. But we also have to think about the feasibility.

And if there isn't a billing code for this, then that means the chart abstraction needs to be done in order to include this variable in a risk adjustment model. And we can talk about whether – the burden that places on providers. But, in the AHRQ model, would it really improve the model so much? I mean, getting from 0.89 to 1.0 is a very difficult proposition. And so, I question the added burden of the NIHSS on the AHRQ model.

With respect to the CMS models, it may have some ability to bump us up from a 0.6 for readmission to something higher, or I think there's a 0.75 for mortality to something higher. But I think for AHRQ, my feeling is that while this might be useful information from a clinical standpoint, I think that the burden of data collection would not be offset by the potential improved discriminative ability.

David Tirschwell: So, thank you for that Risha, and I think it's great to talk about them one at a time. And I guess, I have a – I'll comment that I have a little bit of hesitation at just living and dying by the c-statistics alone. I don't know – when we were at the committee meeting, there was this question of the APR-DRGs being this black box that we didn't know what was in.

And they gave us a website, and I know you had particularly expressed an interest. Did you have a chance to look at the details of their risk adjustment model?

Risha Gidwani: I did have a chance to look at it. I don't – I confess that I won't remember all the details of APR-DRGs now, I looked at that a few months ago. But I think from my perspective, I'm comfortable with the AHRQ inpatient mortality model.

David Tirschwell: And I guess just putting on my stroke neurologist had for a moment as opposed to the objective committee chair – a c-statistic of 0.89 is a remarkably high predictive value. And, that I can't come away from their description of their models with a comfort level that, yes, that mean they're capturing so much data that they can really do this – that it makes sense.

I mean, I guess, I just don't understand where the 0.89 is coming from. And the issue of sort of face validity that this model is really capturing so much of the information that predicts, this inpatient mortality I'm just left a little bit uncomfortable with that. I'd be interested with the opinion of other committee members.

Daniel Labovitz: Daniel Labovitz here again. I completely share that anxiety. It seems, frankly, unbelievable. If we – I've never seen a model be that incredibly predictive, unless you – there are a couple of ways to achieve it. One, you can over model and come up with 8 billion predictors and you put them all into your current sample, and yes, you'll get a high c-statistic. But then, if you try to reproduce it in another confirmatory data set, it falls apart – so called, over modeling.

Maybe, that happened here. I don't know – I don't have a strong understanding of how this was – how the c-statistic was generated then justified. I (inaudible) I don't remember if they did that particular approach of having a model data set and then, a test data set. But I – it just doesn't make sense to me. I can't understand how they managed to get such high predictive value when I at the bedside don't even come close.

David Tirschwell: Yes, I mean, that – I think that summarizes it very well. And I guess, I'm going to propose a specific question if AHRQ has come on the line. How can you help the clinicians that are taking of these stroke patients understand how we can be comfortable with this high level of predictive ability without rich clinical detail, just based on the administrative data?

Is it this just including so many variable that you can predict anything? What is the c-statistic the same in a valid – totally separate validation cohort? I kind of remember that, it might have been. Is AHRQ available?

Patrick Romano: Hi, yes we are. This is Patrick Romano. I'm not sure – am I joined by Jeff Geppert today?

David Tirschwell: Sounds like you're on your own, Patrick.

Patrick Romano: OK.

(Pat) Patrick, it's (Pat) I'm also on the line.

Patrick Romano: OK. Thank you. Yes, so, I just want to point out a couple of things. So one is that, it – there was concern that these statistics might be insulated by the fact that the original model included both ischemic and hemorrhagic strokes.

So, we did proceed with the planned stratification indicator with the development of a stratified model limited to ischemic stroke. And that model had a c-statistic of 0.866 which is a bit lower than what was originally reported, but very similar to what Fonarow reported using the NIH Stroke Scale.

We also looked at where there was some concern raised that the model includes some procedure related variables, such as whether a patient had a craniotomy related to their stroke. And so, we also excluded all the procedure related (intervals) from the model and the c-statistic dropped a bit further to 0.858 – very close to the – slightly less than the value that Fonarow reported using the NIH Stroke Scale.

So, our (inaudible) hypothesis that were testing (inaudible) the data that we used, the HCUP, the administrative data which are the hospital – the all-payer hospital discharge data sets, generally allowed 25 or more diagnosis fields. So, it's much richer than the Medicare data, which until recently have only allowed nine secondary diagnosis.

Second is that the AHRQ indicator is based on all ages and, of course, that allows us to include age as a powerful predictor in the model. And the third is, as previously mentioned, these proxy markers of stroke severity including coma, other alteration of consciousness, convulsions and hemiplegia. So, we certainly recognize that these are crude markers, but in (inaudible) they appear to work nearly as well as the NIH Stroke Scale.

Now, I should also say that we have had a number of discussions with Gregg Fonarow's group and, in fact, we have undertaken – just begun to undertake a collaborative project where we'll actually side by side capture the AHRQ model and to get with the guideline (inaudible). So, a year later, hopefully, we'll have detailed information about (inaudible) the models appear to have similar performance.

I would say if you go back and look (inaudible) but it's not overfitting, because the c-statistic is estimated to have based on a set aside sample, and because the number of parameters in the model is actually fairly (inaudible). So, I don't have in front of me in the order (inaudible) –

David Tirschwell: Patrick, you're cutting out, are you on a cell phone?

Patrick Romano: I'm sorry, sir?

David Tirschwell: You're cutting out. Is that just me, or is everybody else hearing him cut out too?

Male: I'm hearing the same thing.

Male: I hear it.

Male: Yes, he's cutting out.

David Tirschwell: Yes, I don't know if you have a different landline, you could get to that. It might be helpful. So, OK, so you're saying, it was a separate cohort that allowed the c-statistic, so there's no – overfitting shouldn't be a big problem.

And you feel that despite the fact that the NIH Stroke Scale is not apparent, there's enough other information in these 25 other secondary diagnoses that explains somehow more of the clinical outcome possibility than clinicians can really even understand.

I guess, it's sort of that face validity thing that I haven't heard (inaudible) that this huge data set with lots of variables and, of course, the APR-DRG black box. I mean, how many variables do they use? Off hand, my guess is that there's probably thousands.

Patrick Romano: Well, what they do is to classify diagnoses – and it is a bit of a complicated scheme, but they classify a diagnoses based on clinical logic into severity groups. And so, they end up with four severity groups, as we described in the response memo – minor, moderate, major and extreme.

So, everything gets reduced to those four groups, but it's a – it really is a product of a process where a clinician make determinations about which diagnoses represents really severe strokes and which ones represent milder strokes.

David Tirschwell: But they use, you know, all –

David Knowlton: Thousands of DRG – thousands of ICD-9 codes or hundreds.

David Tirschwell: I imagine they use – there's probably some sort of value for almost every ICD-9 code that goes into their black box determination of that final of the four categories.

Patrick Romano: Right. But the overfitting problem again is a problem of over specification of a risk adjustment model. So, in this case, it's not a problem of over-specification of the model. It's just a question of using the richness of information that is contained in ICD-9-CM.

Just as if you were constructing a model based on laboratory data that would potentially come from a pool of thousands of different laboratory tests, but you would select based on clinical logic the specific laboratory test, your radiologic finding that would seem most relevant to the outcome of interest.

David Tirschwell: And, of course, the clinicians would say that those are clinical variables which just aren't available. You know, the other issue for me at least related to these ICD-9 codes is that how they're coded and what's included to a significant degree is based on billing optimization. And as we proceed in the world, and billing practices changes and especially if CMS starts incentivizing or disincentivizing certain types of outcomes, my guess is that the use of these ICD-9 and 10 codes will change.

And so, the whole sort of premise of the predictability might be called into question as the coding changes, whereas if it was a model based on more clinically consistent data that's irrespective of billing issues that would stand the test of time better, and certainly appeal on a more intuitive basis.

What – what's your feeling about that? And I guess, I saw somewhere you annually review this. And, do you change the model every year?

Patrick Romano: The model is changed. It's respecified each year and we do consider user feedback in terms of testing additional variables that should be considered in the risk adjustment. I would say that – obviously, I'm a clinician, so I have, like you, some preference for clinical measures.

But the interesting point is that any data set is based on certain building blocks and is auditable. And so, we have – I've worked extensively with registry data here in California on bypass surgery. And those data have to be audited, because hospitals will gain them in some cases and try to exaggerate severity of illness.

So, this is not a unique problem to an ICD-9-CM environment. It is necessary that whether registry data are used or whether ICD-9-CM coded data are used that the data be subject periodically view in auditing. And obviously, CMS has instituted a robust process as most of us in the industry know for catching hospitals that are engaged in this over coding business.

Daniel Labovitz: Daniel Labovitz here again. I'm sorry to speak up so much, but one of the – something that makes me uneasy about that approach is that we're forever chasing – kind of chase what's accurate. So it's – in taking care of patients with diabetes, we've learned not to give insulin injections based on their last glucose.

But to come with up with a model that works and put them out – on that everyday not to always be just forever giving extra injections where we come up short. And this sounds – this approach sounds like we're always a little bit behind, always – we put in the best model for last year and hope it applies next year.

And, in the finance world that is considered to be – that's what economists do. But it makes very uneasy in the medical world, and past results don't predict future performance is my deep concern here.

I also want to not forget – and I think this has to be part of our discussion, but it may be not right now – the Fonarow papers point that adding in better clinical data – that is the NIH Stroke Scale – changed the ranking of hospitals. So, I don't know how to fit that together. If they're getting – there's a lot to lose there being a hospital for doing poorly – and it would've done well a different model or in a different year it's just not fair.

And I bring out one last thing it's related to that and that is the publication. Neurology came out a month or two ago – I forget – it might have been stroke – (Kleindorfer) company from the Northern Kentucky, greater Cincinnati group, found that the neighborhood you live in predicts the severity of your stroke. And a socioeconomic status is associated – low socioeconomic status is with worse stroke severity and worse outcomes.

Hospitals that deliver care to poor communities may be really hurt by a model like this which doesn't account for that at all. The patients are more sever. They are more likely to die, more likely to get readmitted. The hospitals look bad, but they're actually doing well.

And eager to know the developers comment on that or anybody else's on that aspect.

David Tirschwell: Do you want to ask the AHRQ specific questions, Danny?

Daniel Labovitz: Yes. I guess one like – two questions, one is, how is it – when the Fonarow papers show adding the NIH Stroke Scale changes the ranking of hospitals, can we justify ranking hospitals without it? And two, does the model account for socioeconomic status, and could it be by failing to account for it that hospitals in poor communities will get unfairly dinged?

David Tirschwell :Patrick, are you out there?

Patrick Romano: Yes. Yes. So to answer the first question – so this is an empirically testable hypothesis. The Fonarow's group clearly showed that comparing two models that had substantially different c-statistics, one with the NIH Stroke Scale, one without that the ranking of hospital performance is changed. Would it change to the same extent if you compare two models with very similar c-statistics, very similar discriminatory performance? Maybe it would, maybe it wouldn't.

So that's empirically testable question, and AHRQ has committed the resources to work with the AHA group specifically addressed their question, which we'll do over the next year. But we can't presume the answer to that question in advance.

The other thing – we're not discussing the CMS measure specifically here, but I do note that I agree with some of the responses in the CMS response about the specific way in which Fonarow and colleagues described this problem and reclassification that really may not be

the best approach from standpoint of the users of our measures. But, I'll let them discuss that in more detail when they're up.

In terms of socioeconomic status – it's an excellent one point – we, like CMS, generally avoid including direct markers of socioeconomic status in the risk adjustment models, because we want to expose if hospitals in poor communities are performing worse that that is something that should be understood and recognized by stakeholders. So, there's a danger of obscuring two differences in performance by excusing any difference that you find and blaming it on the socioeconomic status of the hospital's patient.

Nonetheless, if (inaudible) our first findings are interesting and important, we will follow up with some analysis. We haven't had time to do that, but we will follow up with analysis to test whether the AHRQ measure does show some bios against (inaudible) hospitals. I'm afraid I can't address that today.

David Tirschwell: Yes. Well, thank you very much. Those have been great answers and it sounds like you are preparing to do some validation of your model and comparison with the guidelines data set, which I guess, again, putting on my clinician and epidemiologist had I would wonder if that shouldn't be done before this is endorsed as a measure for rating hospitals in a public fashion to make sure – you said we can presume what the outcome of that comparison will be.

But God forbid, I guess it should suggest that there's some misclassification of hospital performance compared to a data set with richer clinical information. I guess that would be a very distressing finding if a measure had already been endorsed.

I would also urge you in this side by side comparison with the guidelines data set to do some very detailed analysis of hospital and/or patient level characteristics that are associated with better than average or worse than average performance.

As you suggest, Dr. Romano that, if socioeconomic status somehow is a predictor of poor performance is that – is that because they're coming in with more severe strokes, or is it because the hospitals that they have access to are providing poorer care. I think that would be very important.

And then one final thing that wasn't in the paper recently in JAMA, but was a question I have at the end of having read it was, what types of hospitals were being inappropriately classified, the ones that were moving from above to average, or from below to average or from average to above?

Were there any consisting characteristics of those hospitals that led to their being misclassified in the first place that might help us understand some of the differences between these different models and their approaches to predicting outcomes?

Any other comments from the committee, especially as related to AHRQ measure? And then, I think maybe we should we focus for a couple minutes on the CMS-Yale measures.

Female: This is (inaudible). I am a nurse and I work with Patrick Romano on the AHRQ measure. I just want to say is the comment about the NIH Stroke Scale that – that is not a perfect stroke measure. And as you all know, that in overrates anterior circulation strokes, left hemispheric

strokes over right, it is not very good on posterior circulation in brainstem. And there are a lot of hospitals that do not use the HIH Stroke Scale.

David Tirschwell: OK. Thank you for that. And I'll just remind developers that unless they are asked a specific question they probably shouldn't comment.

Female: OK, I'm sorry.

David Tirschwell: Anybody else on the committee with comments?

OK. And I'm going to go back to Risha for a second. If you are still out there, would you be willing to comment on at least one – the first CMS measure and whether any of this new information has changed your mind about it?

Risha Gidwani: Well, are you talking about with respect to the AHRQ measure, or to all three of them?

David Tirschwell: I was done with the AHRQ. If you have more comments, please go for it otherwise I'd say we move on to the RSMR, the mortality measure from CMS.

Risha Gidwani: OK, sure. You know, from my perspective as a health services researcher and from more statistical training is that I think, as someone else alluded, it's going to be very difficult to get above a c-statistic of 0.89. I understand some of the clinicians are concerned about the use of only administrative data and the use of only relying on statistics.

From my expertise, the statistics in the model are really all I can speak to, and I am comfortable with AHRQ model as is without the NIHSS. I think as we move to the CMS model, their model diagnostics indicating some opportunities for improvement in this coming visibility could benefit from NIHSS.

David Tirschwell: OK. Do you have any specific questions for CMS-Yale – the developers – about their models especially as it relates to this, the new paper that was published in such?

Risha Gidwani: (Stephen) my questions actually related back to the original model. And I think I brought this up in the meeting, and I still feel as though I am not entirely clear on the response. But when I look at the model diagnostics for the readmission model, it shows that cerebral hemorrhage and hemiplegia, cardioplegia, paralysis or functional disability are protective against readmission. And I'm afraid I thought that has for face validity for me, and I'd appreciate some greater insight into how this could be.

David Tirschwell: Thank you. That's great. Would CMS and/or the Yale group be willing to respond to that?

Susannah Bernheim: Yes. Hi, this is Susannah Bernheim from the Yale team and I'm happy to. It actually is a great opportunity to comment on some of these issues.

So I'm going to start with the particular issue that I believe that was Risha had just raised which has to do with the c-statistic. And I want to make a statement that you made earlier and just echo it which is the importance in differentiating between a model or your trying to best predict for patients, what their outcome is going to be in a hospital model.

And a hospital model that's trying to risk adjust to level the playing field is choosing which adjustment variable very specifically that were inherent to that patient prior to or at the time of the admission.

So when you look at our model and the AHRQ model what you need to realize that they're both faced on the exact same ICD-9 – the AHRQ model has more of them, but is changing in the CMS measure. There's more during the inpatient. Our model has the ability to look back 12 months and get any outpatient or inpatient ICD-9 codes.

We have a pretty rich data source for the patient, but it's all administered claims. What would potentially give our model a much higher c-statistic is if we risk adjust it for things that happen to the patient after they arrived in the hospital, and we're very careful not to do that.

The risk of doing that is that you then give hospital essentially credit before quality and put them on a playing field that doesn't reflect their quality. So one way to increase the c-statistic and get good at predicting whether a patient is going to die is to risk adjust for factors that happen right before they die.

David Tirschwell: You're not suggesting that the AHRQ model is doing that, are you?

Susannah Bernheim: I'm not, but I'm trying to clarify about the (inaudible). So if we are risk adjusting for things that might be outcomes of a stroke a patient came into the hospital for, we are not risk adjusting for them during the in-depth admission. So those represent prior claims. Those represent patients who had history of such things.

So that was – the first – it is a caveat against that c-statistic issue and looking for too high a c-statistic because what you can often find in that case is inadvertently we are adjusting for things that happened after admission.

The next important – I'm, sorry – pardon?

David Tirschwell: I was just going to say that I agree. We don't know when that hemiplegia referred to. I guess I assumed that AHRQ is, at least going forward, going to be (applying) that code to a present on admission thing, but that's fine. I guess I had to ask you to try to specifically respond to Risha's questions.

Susannah Bernheim: Right, so I'm trying to. I think there's a couple of – but it needs to be – I think it needs to be in the context of what this measure is trying to do. So the other important thing has to do with the measure representing a surrogate for the clinical data that one would like to have. And the problem with the clinical data, as other people would have said is that even in hospital that have enrolled in a registry less than half the patients who have NIH Stroke Scale.

So what we have to ask ourselves is, are we able to adequately risk adjust? Are we able to adequately understand the risk of the patient of the hospital population without having access to that data, because we don't feel comfortable leading half the patients out of the measures?

And we have three studies that have looked at this. Our own analysis – or I'll remind you that in our own analysis we had a measure stroke severity on all of the patients that we're including in a match cohort and had a high correlation, the recent (VA) analysis which is in a

different population, and in the recent – with the guidelines paper. And in that paper we don't have information on 55% of the patients.

The model that's used is not our model and doesn't account for this (ED) transfer patient which is an important way that we're accounting for some of the severity differences. And as we discussed, there's not a – the reclassification analysis is not done with an understanding of the full case mix of a hospital.

David Tirschwell: OK.

Risha Gidwani: (This is) Risha. And so, in terms of the hemiplegia question, are you indicating that the hemiplegia would be something that occurred to the patient after they were admitted to the hospital as part of poor clinical care in that (inaudible).

Male: I think – this is (inaudible). Let me just try to explain, because we've had this encountered some times variables that go in directions that are maybe counter intuitive. For example in the MI and heart failure, hypertension which would seem the important comorbidity actually is protective against mortality and in some cases protective against readmission. You know, how these things play out often difficult to know and the empiric result by the way which can be bare out with the clinical data too don't know which go in the exact directions you want.

What's important in the way that we've done is focus our attention on the result of the administrative model and to say to what extent does the result the output of that model, the characterization of a hospital's performance corresponds to the output of a model that's built based on really detailed clinical information including a stroke severity scale which is considered to stand very well to the NIH score that was available on all the patient.

And so rather than dig into each of the variable exactly correspond to what's in the chart about that variable. This is about thing that when you put all these variables in, does the output are event in measure development? Because when I first started this, I was really against administrative claims. I said you'll never be able to convince me that we should be using this for public reporting.

In the breakthrough was the thought that actually I don't care about the individual elements. What I care is that the output of the model is does it represent a reasonable surrogate for the output of a model that has a higher c statistic in fact and is more rich and I could use to predict individual patient outcomes much better. But it's the hospital level does it give me a result that's very close.

And in the case of M.I., heart failure, pneumonia and now stroke, what we find is the use of a mystery claim found the way that we do with just the data that would be available on admission lines up very well with the output of a model based on clinical data, which by the way has a higher c statistic. But at the hospital level, the characterization of performance is very close.

And so then we're immediately (part) of the article. He did not do that. He also didn't compare our model. He also compared only 45 percent of the patients in their entire sample. And he is using NIH Stroke Scale, which by the way is very considerably across the hospital, so we're not even sure that the degree to which it is accurately representing the patients.

And I can tell you working with registries where they're voluntarily inputting data and there's little oversight and quality control over the type of data that they input you can't be confident about it either. So…

Male: I was going to – let me just interrupt for a second, because we're straying a little bit from the question at hand. You've now referred a couple of times to a comparison of your models to a rich clinical data set. And was that part of the neurology publication about the RSMR? Because I just don't remember that off at the top of my head.

Male: Well that was part of our application where we – when we can, this is our usual approach which is we get a data set, a broad data set where we can get detailed clinical information about the patient's case on chart obstruction. Where if you had enough money and we had perfectly (inaudible) around the century this is what you want to do, which is you get vital signs and clinical status and this now came from the National Stroke Project that CMS implemented across almost 30,000 people represented a sample across the United States, where, you know, I think 200 or 250 elements were obstructive from the chart.

And we created our very best clinical model which included a stroke scale that again since nobody is collecting the NIHSS in a regular fashion we didn't use that. But we used a stroke scale that has been – there's an article that says it stands quite well and NIH Stroke Scale – NIH Stroke Scale is slightly better but they are pretty close.

Male/Male: (inaudible).

Male: What's that?

Male/Female: (inaudible).

Female: …looks at functional deficit in four different ways. I don't have the name on top of my head. I can send that…

Male: We can send that, too. But like she said it has functional deficits and we were able to collect it on all the charts and in this inner application. And then we created profiles in this case of larger geographical in the hospitals, because that was the way the sapling went. And we said to what extend does an administrative claims model then produce output, a performance that would (inaudible) the chief physical and clinical model which did not include any complications.

Patrick would have to say whether or not they're 0.89 includes potentially complication. But we don't use complications with a very rich clinical data set. We get about 0.8 and that output characterizing performance (inaudible). Now if I use administrative claims bonded, the best job possible how would the output, the characterization of performance compare because that's what we care about at the end of the day which is whether mystery claims model can service a surrogate for the model we really would prefer, which is the model that was reaching clinical data.

And what we found time and time again is the mystery claim though actually produces output at the organizational level that's very comparable, very similar. And that's we got knee over the hump to say, you know, from fighting this, because when I first start saying we do this for same as that, it actually took the contract to show them why they shouldn't be

doing this. And it was really this insight about – this is all about – does the output of the one model, how much does it agree with the output of a model that I prefer to be using.

Male: And what was the answer to that? What was the correlation in the rating of individual regions or hospitals?

Male: The correlation was 0.8. Again, we can show you the thoughts. It's all in our application. But this was I think the critical nature of this, because then it gets you away. That means this is where the c statistic falls, because you compare two models with very different c statistics.

But at the institution level, they're producing very secure, high on one, you're high on the other. If you're low on one, you're low on the other. And there's, you know, relatively little play. And the question for the committee probably is how good is good enough for that agreement. Because if that agreement is not good enough then it's a good reason to say then this doesn't work. If the agreement does seems good enough then it doesn't when we argue about individual verbal or (inaudible) I think we're removing off the main question that should be focused on when considering whether the model is good enough for profiling performance on a public basis.

Male: Yes. Are there any standards to present good enough?

Male: Yes. That I guess, you know, you guys can establish a standard. It's not – because neither measure is going to be perfect, because there is a possibility of some game we've noticed take for example shock and heart failure. I'm a cardiologist so I'm much more familiar with this.

We put those in our clinical models, but the variation of cross institution is hair raising. I mean in some institutions, 40 to 50 percent of patients are called having shock and some institutions it's 2 to 3 percent and we don't thinks that's because the patients are so different.

We've just developed e-measures and we basically have banned the use of things like shock in the e-measures, because we don't trust them enough to be consistently documented. And so the definitions aren't being used by the doctrine in a standardized way and we've abandoned them.

So even the clinical model is not perfect, but, you know, whether it was good enough agreement between these two (incorporate) models, one having a lot of the data that we would like to have because as clinicians we tend to at least trust the vital signs and severity scores more than we did the administrative codes, how it's good enough is still yet to be – yet to be determined. And this is I think something we should always struggling with.

Male: I agree. And the follow-up question, what's, I mean, I take away from here already the things that you've said that you have some issues with the paper by Fonarow and the Get with the Guidelines group. But specifically, I mean, I guess you all in supportive of the measure that's being put forward, which suggests that the reclassification that they highlight, the hospitals change categories, you know, how do you – how do you get us around worrying about that?

Male: Well, I think there are a couple of things I wish (spread work) with us. We went to (DHA) a year ago and (Greg) you might be on the front you remember the conversation. I was trying to broker this and we didn't quite. I have a lot of respect for (Greg) and I think that their

contributions are useful in terms of raising hypothesis about future work, but I don't know about its relevance to the exact measure, because they didn't do this measure.

And I terms of the reclassification, if you create strict categories and I know from a policy point of view this is what happens, but I've been arguing about it from a policy perspective, too. There is uncertainty associated with each estimate. So even just pass retest there's some slight movement.

Most of the hospitals did stay within the same category, but I think a better way to do would have been to show the continuous variables and to say, "How much did they agree with each other," in other words, same way that we did it model produces. This is similar with this model. These are similar with the other model. How close are they when you create like, "Do they cross arbitrary boundaries where there are a lot of hospitals that are right on that boundary with one on the other might just move slightly."

And whether or not that's important or not is it depends on the policy. Like is aid I've been already against the national policy that takes the national, I mean, I'll just say it, I mean, (inaudible) is laughing at me. But I think there's uncertainty such that estimate and hospitals are right next to each other from a payment policy can be penalized in different ways and I don't like it. I don't think it's right, I don't think it reflects the science.

But, you know, with regard to measurement, I think the best thing we can do is just reflect the best science. So I think great paper is very helpful in terms of thinking about the next steps. I didn't think it would do everything a verdict on this measure, because they weren't using our measure and I didn't think that the way that they did it was the way it should have been done with respect to validating and demonstrating servicing, which is according to the way there are methodologies that have be used I would have like to see now.

Male: OK, thank you. Thank you very much for those responses. Does any of the steering committee have any further questions for the CMS-Yale group measure developers?

Risha Gidwani: I have a comment. This is Risha, good morning again.

So I'm just reviewing the information provided originally by the developers. And looking at the medical record model versus administrative data and showing that they have this correlation of 0.80, although it does show that the mortality medical record model, the data that was – the model populated with chart obstruction had a higher c statistic of 0.80 than the administrative data.

But even if they have a high correlation of 0.80 between the chart obstruction and the administrative data, I think the question is that the Fonarow paper is raising the issue that whatever administrative – whatever model you use adding an NIHSS Stroke Scale may improve that model discriminative ability. Even though they use a different model in the Fonarow paper than the CMS paper, it seems to me that the question we're talking about here is not changing the entire CMS model but rather just looking at what would the CMS model discriminative ability look like with and without the NIHSS included.

Is that something that could actually be assessed so that we can get an understanding of what this new information would do in terms of predictive ability?

Male: I think first of all that was inept because it was the same model plus or minus the variable which has always been an improvement to its performance. So I give you – you give me (inaudible) thing variable and I add another variable as any predictive quality has been reduced to variant split.

The problem is with all due respect with the guidelines, you would have to impute the NIHSS for 55 percent of the patients in order to really have the full sample to know that you are confidently evaluating the hospital's full experience. You're evaluating a certain proportion less than half on average of the hospital's patients when you're making this comparison. And I don't know what that does to the evaluation whether or not advising is one way or the other or whether it's neutral.

And so I think that by itself not only telling you that only less than half of patients are having these variable collectives, but it's a great impediment to doing the validation because I validating against non-randomly selected subset of patients that varies across the institution. Some institutions are having more NIHSS scores obtained from having less and I think that I wish that there was a simple fix.

Patrick can do this with Get with the Guidelines, but I think that there's severe limitation and unfortunate. And maybe what we need to do is get people collective more. But for right now, we have, you know, we maybe could get (DHA) to come to the table now and do this. But I'm still not sure what it would tell us with 55 percent of the patients having to be excluded because they don't have the score.

~~Female~~ Risha Gidwani: May I just – I just like to bring into specific two points there. One is about including another variable (that tell us) the variation, ~~including another variable in the risk reduction model will increase our score because the extreme amount of variation in your outcome that is displaying by your predictor. But we're talking about here is the c statistic with the discriminative ability. So I would like to distinguish between the improved ability to exact variation and the important in the c statistic, which are the same.~~

At the request of Dr. Gidwani, this comment has been amended to more accurately reflect her statement: "Including another variable in the model will improve the R squared statistic because in adding another predictor you increase the amount of variation in your outcome that is explained by all of your predictors. But the C-statistic is different and tells us about discriminative ability. So I would like to distinguish between improved ability to measure variation and discriminative ability, which are different things."

And the other suggestion that I might have is that understanding that only 50 percent of the sample would have NIHSS recorded, it maybe possible than to just look at the subset of patients that didn't have NIHSS and then run the administrative model and test those patients one model without that NIHSS and then one model with it.

Male: Right. I'm just thinking you just don't know where that's going (bite), because if not a 50 percent across all the hospitals it's the variable amount of missing across institutions and what impact that has I don't know.

Well, you're right. You can do it. I'm just thinking it will still is open a lot of questions, because it's more than half are missing and it's just unfortunate.

Daniel Labovitz: Daniel Labovitz here. Looking at this from slightly broader perspective, I think we've talked a lot about the NIH Stroke Scale as a flawed scale. It doesn't substitute for real neurological exam. And yes, it's missing in a huge percentage of cases in the Get with the Guidelines data set. But what was shown in that data set flawed as it was is that adding that variable in made a difference and it improved the models substantially and it changed the ranking of hospitals.

If the fact that NIH Stroke Scale got problems and the fact that there maybe bios, I don't think changes that concern about leaving the variable out. It introduces the real valid question, is the administrative data truly enough. And I'm still – I remained, you know, I think the discussions so far has been compelling and I'm really glad that hearing it on the phone really helps me understand how it is that a cardiologist in (inaudible) say, "Yes, gosh. The administrative data really does work." I really respect that.

But I'm still deeply troubled that we may be moving to this a little too quickly, because we haven't done – we haven't done enough background to work. This needs to be sorted out, as Dr. Tirschwell said this needs to be sorted out before we approve the measure not afterward.

(Dave): Thank you, Daniel. Any other comments from committee members about these severity issues?

Bob Barsan: Dave, this is Bob Barsan. I think I came into this discussion with the same bios that the Daniel had. I guess I'm feeling, I mean, it has been a great discussion I think I understand a lot of the issues a lot better. And I guess I'm less concerned than Daniel that this data while imperfect maybe worth than having no data.

I mean I think I've been convinced that it's probably reasonable data could it be better with adding the severity indicator like NIHSS, you know, like the answer is probably yes. It's good that they're working on that together and hopefully in the future we'll have that. But I guess I'm less concerned that this might that this might be highly inaccurate, you know, the way it is.

Female/Male: (inaudible).

Male: …object – oh, sorry. Go ahead.

Female: No, I was – this is NQF. Dr. (Schwam) has requested to speak to the paper. Is that all right with the co-chairs?

Male: It's fine with me.

Male: It's OK with me. That's fine.

Dr. (Schwam): Oh, thank you. I just wanted to make the committee aware of the fact that at the office of the JAMA paper, we are on the call and there have been lots of characterizations of what the paper did and didn't do. If it would be helpful to you, I would be delighted to Dr. Fonarow a minute to just maybe address some of those of concerns.

And I think, you know, the overall context that I would echo is what one of the committee members said which is the impact of including the score whatever the settled differences are

in the models. And I think you'll get many different opinions about those differences adding a clinical variable which is present on admission and it's easily performed.

It's not a question of this being a difficult test to administer. The reasons there are 50 percent of the patients having it in the Get with the Guidelines is that those hospitals made a commitment to collect it. Hospitals don't collect it because they're not paid to collect it and they're not required to collect it. And if we require them to collect it, they could collect it tomorrow.

So I just think again I'm concerned about dismissing the added information contained in this very important clinical variable. But let me turn it over to Gregg to just address any aspects of the paper that you think need to be clarified.

Gregg Fonarow: So I would just tab to what we provided is, you know, we did make our very best effort to approximate the Yale-CMS model using all publicly available information. I know one covariate did subsequently added but all of the other variables were in common in the performance of the administrative model and our paper is very similar to what was provided to use. So we really feel with any administrative claims data using pre-admission variables that the addition of the NIH Stroke Scale would allow hospitals to be more accurately classified.

This is a unique variable that in every paper that is looked out at the (peep) shouldn't level in the hospital level dramatically improve the discrimination and performance of the model and that ultimately translates into meaningful reclassification of hospitals. So I do think that potentially having a public reporting hospital is having worse or better than expected mortality that's not actually accurately and reflective with their true performance can indeed harm stroke patients in those hospitals that care for patients with greater stroke severity.

Male: can I ask a follow-up question, gentlemen? Thank you for being on the call.

Dr. (Kromo) I think appropriately brought up the issue that with every data set of patients that have stroke, there will be some drift. Some hospitals will go into a higher category out of a lower category that's just inherent in the data and the variability of data. And of course you point out that with the non-NIHSS model versus with some go in and go out. Is the amount – did you look to see whether the amount of sort of natural drift in and out is different than the drift in and out you saw with the addition of the NIH Stroke Scale?

Gregg Fonarow: So we've been interpreting that question in the issue of being able to categorize the hospitals using different methods that those categories changed dramatically when you further adjust for the NIH Stroke Scale. We find that at the hospital level that there is variability in the severity of stroke among hospitals and those hospitals that were having patients that on average had more severe strokes in those hospitals that on average had patients with less severe stroke and that had the greatest meaningful impact on their relative performance ranking is well within absolute terms how their risk standardized mortality rates differ between the two models.

Male: And just to amplify, I think, you know, Bill Barsan raised the issue of, you know, is it – is it really going to make a difference sort or is it closed enough. I think the biggest concern I have as a stroke neurologist is if hospitals who admit patients or receiving transfer patient with higher NIH Stroke Scale scores or get them directly, because in their city there is

(inaudible) algorithm such that more severe strokes go to stroke centers and very severe strokes go to comprehensive centers.

The centers that receive the patients with the highest NIH Stroke Scale scores will have the highest mortality. And if those publicly reported data put pressure on hospitals either through changes in reimbursement or changes in reputation, they will have an incentive to discourage transfer or acceptance of the sickest patient. And I don't think any of us want to see that happen and that's my biggest concern, not the gaming and not the, you know, misclassification within categories that are meaningless.

But the real danger that hospitals that currently are the safety nets and take care of and admit the sickest patient and support them through their decisions to enter hospice and die in hospital or shortly after discharge will have a barrier to continuing to provide that service.

Male: This is (inaudible). Can I provide one point of clarification?

Male: Sure, go ahead.

Male: I just want to say that I think we point here is a very important one that we listen very carefully, too. I want to show you how the process within NQF of approval goes. He brought this up to us that is one of our members of our technical advisory panel that the patient to some of these hospitals might be getting transfers in. And these transfers in it's true we validated this tend to be sicker and which is contradiction to cardiology where they see they're healthier. But in neurology, they are sicker and that those hospitals might be then putting a position of looking worse.

So this was the variable that we created, transfer from an emergency department to the hospital that became a very important variable. This is the variable that (grades) there was a minimal variable, but in fact we believe it to be very important one. We congratulate and thanks (Liz) for bringing it up to us.

And this is the variable that was missing in the comparison in the JAMA article and why we think that in the end you couldn't really compare them because we responded to that public comment, our technical adviser panel comment. We improved our model and that we think that that made it even better than it was before.

So the question is now, you know, could it be good enough? These are all the issues, you know. Are these good enough? But I just want to at least say we responded to (Liz) very thoughtful comments that these transfer ins might be causing (inaudible) that is in the model whether the rich patient who is transferred in from an emergency department, not from a hospital but from an emergency department to a hospital where they're subsequently admitted for stroke that is now a severity measure in fact it's an important one. We were interested it's a strong predictor about coming out incorporated.

(Dave): Thank you for those clarifications. I did want to just add one thing that wasn't captured in the summary memo from NQF that struck me in the comment period as well as they did today in the in-person meeting. And that is that the pretty much all of the large provider groups that take care of stroke patients have come out against these three measures due to the lack of severity adjustment.

And so there were comments from the American Stroke Association, the American Academy of Neurology, the American Academy of neurological Surgeons, and the CDC which runs the Paul Coverdell Stroke Registry all very worried about these measures without severity adjustment. And so those are – in my mind those are our big players and represent significant stakeholders here.

Any other comments that people have about all of these mortality measures or questions for the developers before we move on?

OK. Given that, the next item on the agenda has to do with additional discussion on other comments and responses that we didn't already include. I already did that in some degree. Anybody else have a comment that they wanted to bring up that we haven't already discussed?

Dave Knowlton: Dave, it's Dave Knowlton. I just had a brief comment. I didn't know the proper place for it.

(Dave): Go for it.

Dave Knowlton: That the, you know, I hear all the discussions and a little bit (inaudible) to a non-clinician. But I on the one hand want some robust severity adjustments so that we can trust the measures. But really from the purchaser-consumer standpoint, which I represent that's my constituency, trying to get some measure that allows people to discern between a very noisy marketplace where everybody is claiming to be able to deliver competent stroke care at the secondary even in tertiary quantity level is a big concern.

And so to the extent that sometimes I just hope that the committee as we look, and I assume we're looking this in greater depth this group, will keep in mind worrying about the perfect leaving out the possible. We (made a stat) getting some robust measures out there.

And so I don't mean that to speak against is very quite a contrary. It's important that we have a decent severity measure so that the (page) is credible. But right now we have a lot of people claiming the capacities they don't have. And so from a purchaser-consumer standpoint that's where the bulk of my concern a lot rests.

(Dave): So you're suggesting that we consider the fact that we shouldn't let perfect get in a way of good enough starting point?

Dave Knowlton: Yes. We need to start getting some way to sort the grant from the job out there. And I hear the pushback we want to do so fairly. Usually in quality measure sets means being as fair as you can be with the data that you currently have. And we would love to have clinical data it's always wonderful, but is not usable in the public domain unless you can get in the greater correlations that are significant.

So there are going to be some real issues here. I'm pleased we're having this great discussion. I hope we'll be able to guide with some more.

Female: And David, this is –

Female/Female: (inaudible).

Risha Gidwani: This is Risha Gidwani. I think those are some excellent points and certainly well worth considering. If I could also just add though that through these measures that are coming from CMS and CMS has instituted a program that will adjust hospital reimbursement based off of a measures MI, heart failure, and pneumonia is not unreasonable to think that the stroke might be next and that also there are also payment adjustments that will start to implemented for readmission.

So I think there are two sorts of things we need to consider. One is this very thoughtful and true point about hat the public health care decision making than the other being also the payment adjustment could very well follow from endorsement of these measures.

Helen Burstin: And, David, this is (Helen). If I could just follow up on that comment; Helen Burstin.

Just in general, we try to keep a line with between the endorsement sides. We were looking at the measure qualities themselves from the potential users. Any endorsed measure could be used for a variety of accountability application as well as quality improvement and it really is another process to measure the application partnership that will help to determine which measures get used in payment.

So it's a slippery slope. I just wanted to try to stay focus on the quality measures at hand fully knowing that if it's an endorsed measure it is potentially the applicable to a variety of those applications.

And then one last comment, Patrick had asked us to just remind the committee, I guess there was misspeaking earlier that the current AHRQ measure on inpatient mortality is already endorsed it's not a new measure. It's up for maintenance.

Karen Johnson: And this is Karen, David and (Dave). If we can just get clarity so that we know what we're doing after this call, has the committee decided that they want to revote on all three of those measures or just on the CMS measures? Can you just make sure that I know the answer to that?

(Dave): Well, that's a great question. I'm not exactly sure how to do this efficiently. Let me – let me just cut to the extreme. I guess I would suggest that we go ahead and revote on all three of these risks adjusted measures. I didn't hear anything to suggest that any other measures needed re-voting. If you have an objection or an alternative suggestion, please pipe up now.

Male: Just a question, (Dave) or Helen. When will we do that?

Helen Burstin: This is Helen. We'll likely prepare the materials the summary of this call. We'll get the transcript and we'll place SurveyMonkey out sometime this week, do within a week or so.

Dave Knowlton: (David), this is Dave. I think I agree with your assessment. I think that's exactly what we should do. But I wanted people have the sense to know what happens. So I'm (inaudible) that somebody else objects to it.

(Dave): I'm not hearing any objections.

Dave Knowlton: Well then let's do that.

(Dave): Great.

Karen Johnson: OK. The project team will get those materials out here as soon as we can. And this is Karen. We could talk about it a little bit more or maybe I could just sit in and send it here and that might be enough.

When you are thinking about the two CMS measures, please take a peek at the latest reports that CMS you all provided. They have re-specified their measures somewhat. Specifically, they have changed their measures to be looking only at pay for service 55 and older. They made a change to where their measures now would be specified for all payers 18 and over, so a pretty big change there.

And then on their readmission measure, they have also made a change in terms of what, how they define plan to readmission. Earlier, when you saw the measure come in the first time around, they had a set of specific measures or a specific list of readmissions for stroke.

What they've done now is harmonize with their other readmission measures that are already in NQF roster in the process of being endorsed by NQF. And it's a bigger list if you will of readmissions that they consider plans. So the exclusion with that measure is now a little bit larger.

So as you are re-voting those measures, please go back and look at the new materials and just make that it continues to satisfy the scientific acceptability criterion.

Susannah Bernheim: And Karen, if I could just clarify for the committee. This is Susannah Bernheim from the Yale team. Those of us has done in response to stakeholder input and the committee's input, do the all payers testing that we did since your last meeting was in response to the request from the committee that we specify this to be an all pair measure aligned with ours and the planned readmission algorithm that's one that the hospital association has asked us to apply the measures and so, so that's why that specification happened at this stage in the process.

Karen Johnson: Thank you.

(Dave): OK. So we've move through the bulk of the agenda. The final bit refers to additional areas per measure development. And there were a number of comments on that. Somebody suggested a better outcome measure would be endpoint of death and severe disability. Of course that would be – that would require additional data collection similar to the issues related to the NIH Stroke Scale but I'm sure all would agree would be superior.

There was another suggestion for a measure to document patient and family training and education in acute and post acute settings to reduce disability burden of care in primary and secondary prevention. Any comments on those new gaps?

There was a suggested edit that measures of post hospital care be changed to, the wording be changed to measures of post acute care and rehabilitation care. And there's a proposed committee response agreeing for the – agreeing to the suggestions for future measure developments. And I think also probably agreeing with the, what seems to me reasonable changing in the wording for the post acute care and rehabilitation care.

Anybody care to comment on that or suggests an alteration to the proposed committee response?

OK. And then the next step on the agenda is a little bit of time for public comment. We probably have about 15 minutes total for this. NQF do you – is there – shall we just ask for public comments? Is there specific (queue) that you've developed or a way for the public to make themselves known that they would like to talk?

Female: The operator – Amy, can you open the line for public comments? And then we'll just take them as they come in.

(Dave): OK.

Operator: At this time, if you would like to ask a question or have a comment, please press star-one on your telephone keypad.

(Joe Broderick): Hello?

Male: OK, somebody hit the one. I heard two of them. Go ahead.

(Joe Broderick): This is (Joe Broderick). I don't know if – can you hear me?

(Dave): Yes.

(Joe Broderick): Great. Well just to introduce myself. I'm Joe Broderick. I'm at the University of Cincinnati. And some of the comments before about the greatest (inaudible) Kentucky Stroke Study that was the study that I began back in long time ago and the comments of that economics were appropriate.

But I really want to speak as clinician. And somebody involved the epidemiology of stroke and also someone who knows how it really reflects what happens here in our community in Cincinnati, which I think is relevant to all those cities across the country.

The EMS here takes patients to all the hospitals. But for a number of these services they take it to the hospital that requires certain severity. So if you have a more severe patient, they will tend almost certainly to go to university. There will be certainly patients that are transferred, so I'm glad that transfer issue has been – is being addressed regardless on what model are you talking about.

But this issue of severity of the stroke patients and how primary stroke centers and comprehensive stroke centers will be the most likely to receive those kind of patients in most communities in United States and what will happen is they will have the most severe patients and those patients will have the highest mortality and it will actually make it look as if the primary stroke centers and others are doing more poorly compared to the other hospitals which received various few stroke patient.

And so as a clinician, I'm just saying empirically that's how patients come to hospitals in this country. Unless you have a way of measuring the severity and there are different ways and I spoke (inaudible) one way of doing it, you're going to have a very imperfect model no matter

however it's going to be set up and you're going to probably misclassified people in hospitals.

And that will be something that I'll be highly not in favor of and I think the various group of the measures need to work together to try and come up with better solutions currently out there that will be like. That's all I need to say.

(Dave): Thank you very much, Dr. (Broderick). Was there another public comment?

(Esley Schwam): David, this (Esley Schwam). Well we will submit our comments to NQF in writing in response to some of the questions that were raised about the statistical methodology and won't take any time on the call to do that.

(Dave): Thank you for doing that. As you probably heard the voting on these measures will be occurring over the next week or so, so to optimize the chance for those materials getting appropriate review, obviously the sooner the better.

(Esley Schwam): Yes. I think we'll make an effort to get them in by the end of the week. I think the only public comment I would make is in response to the gentleman who represent the payers, who said that appropriately the public and payers need information to help them properly choose hospitals that are providing the best care.

And I think (Joe Broderick)'s point emphasizes the fact that misclassification of hospitals actually provides the public with an extremely confusing set of information in which their local community hospital appears to be the provider of the most superior care in their nearby university hospitals perhaps average or low on the list. And that also is confusing piece of information for patients and payers to try to integrate into their understanding of where resources should be directed. Thanks.

(Susannah Bernheim): (David), this is (Susannah Bernheim) from the Yale team. Can I just make one quick comment on the two concerns just to remind the committee of some analysis we've done.

(Dave): Sure.

(Susannah Bernheim):I just want to briefly there has all along with this I think rightfully been a lot of ear about how these measures will characterize academic centers, stroke centers, centers that receive a lot of patients in transfer. And over the tiers this has all been the measure we've looked at that from many different angles. Again every time we've looked at it, we've looked at how stroke centers (inaudible) academic teaching centers we look. We also have hospitals who receive a lot patients to transfer and look and we never see any real difference in addition to the submission on the results or certainly not that those centers are looking worse, sometimes they are looking better.

So and again I would say we are sympathetic this year, but when we have tried very hard to analyze it from every way to see if there's anything that is true with the measures that we produce or not seeing that that is actually how they profile hospital.

(Dave): Thank you. Anyone else with any comments or questions from the public? If not then I guess I would like to turn the meeting back over to NQF to briefly review next steps and then we can adjourn.

Female: Thanks everybody for our great discussion today. I will be sending out the transcript, some notes and a voting survey in the next couple of days. We will get that out to you as soon as we can along with the timeline for when we need your vote by. And then we'll be moving forward to NQF member voting following the conclusion of your voting. For the Steering Committee members that are serving at phase II of the project, they'll be receiving more information about phase II tomorrow. So we look forward to hearing what you think about these measures and I'll get those materials out as soon as we can.

There are no questions. I think we can adjourn.

(Dave): Thank you everybody for your participation.

Female: Great. Thank you, (Dave).

Male: Thank you.

Male: Thank you. Bye-bye.

END