NATIONAL WEB-BASED TELECONFERENCE ON HEALTH IT: QUALITY METRICS AND MEASUREMENT

BEGIN TRANSCRIPT:

MALE MODERATOR: Hello and welcome to the AHRQ Webcast. Today's topic is Health IT: Quality Metrics and Measurement. At this point I'd like to introduce today's moderator, Angela Lavanderos, Program Analyst with the Health IT Portfolio at the Agency for Healthcare Research and Quality. Angela, the floor is all yours.

ANGELA LAVANDEROS: Thanks Colin. Before we begin today's session we are required to read the following statement for CME purposes. This educational activity has been approved by the Wisconsin Medical Society for 1.5 AMA PRA Category One credits. Speakers and planners are required to make disclosure of any relevant financial relationship which may be related to the subject matter discussed.

Dr. Mark Weiner disclosed that he had received grant support from Pfizer, however the planners for this activity have determined that this is not a conflict of interest relevant to today's presentation. All other speakers and planners for this activity have made proper disclosure and have no relevant financial relationships that exist now or in the past twelve months.

So now I'd like to introduce to you the three presenters for today. Dr. David Baker is a Michael A. Gertz Professor of Medicine, Division Chief of General Internal Medicine and Director of the REACH Practice-Based Research Network at the Feinberg School of Medicine, Northwestern University. He has served as a member of the Health Information Technology Expert Panel (HITEP) Quality Data Set Subcommittee, the Physicians Consortium for Performance Improvement Measure Implementation and Evaluation Subcommittee, and the American College of Physicians Performance Measure Technical Advisory Committee. His research has focused on health literacy, racial and ethnic disparities in care, and quality of care for chronic diseases. His current research is examining the use of health information technology for health communication, quality measurement, and quality improvement.

Andrew Hamilton is Chief Operating Officer and Director of Clinical Informatics at the Alliance of Chicago. Mr. Hamilton is a Masters-prepared nurse informaticist with twelve years experience in both inpatient nursing care and outpatient community health, as well as nursing administration. As the Director of Clinical Informatics, Mr. Hamilton developed clinical decision support and national clinical performance measures for the organization's electronic health records and meaningful use efforts.

Currently he serves as the immediate past president of the Board of the Centricity Healthcare User group and is also a member of several local, state, and national EHR and performance measurement related workgroups. In addition, Mr. Hamilton is an adjunct faculty at Loyola University School of Nursing, the University of Illinois at Chicago School of Nursing, and recently joined the faculty at Northwestern University's Medical Informatics program. Previously, Mr. Hamilton was a pediatric critical care nurse and a member of a large academic hospital health IT team, supporting the implementation of clinical information systems. He also served as the director of Patient Care Services for Howard Brown Health Center. Mr. Hamilton holds a B.S. in Nursing and an M.S. in Nursing Business and Health Systems Administration with a focus on Nursing Informatics, both from the University of Michigan School of Nursing.

Dr. Mark Weiner is an Associate Professor of Medicine and a practicing Internist for over 15 years at the University of Pennsylvania School of Medicine. Dr. Wiener's research interests helped to bridge the gap between health services research and medical informatics. Having completed training as both a VA General Medicine Fellow and a National Library of Medicine Fellowship in Applied Informatics, his work helps to create systems to adapt clinical and administrative data for research and quality improvement purposes.

Dr. Weiner has overseen the technical development and uses of the Pennsylvania Integrated Clinical and Administrative Research Database system, the PICARD System. He is also the Director for Information Systems Integration for Research within the Office of Human Research, Co-chief of the Biostatistics and Informatics core of the VA Center for Health Equity Research and Promotion, and Co-chair of the Data Core of the FDA Mini-Sentinel Initiative. His work on the Medical Home Initiative at the University of Pennsylvania Health System helped him to formulate the goals of an HRQ grant to apply routinely collected data with electronic health records to improve quality measurement.

So with that I'd like to begin the teleconference. Dr. Baker will present on the Utilizing Precision Performance Measurement to Improve Quality Project, the UPQUAL Project, which targeted four chronic diseases and five preventive services for physicians to achieve ultra-high levels of performance. He will described the rationale, provide an overview of the intervention, and present the changes in performance after one and three years.

Dr. Baker, the floor is yours.

DR. DAVID BAKER: Thank you very much, Angela. I just want to stay at the start, thank you to AHRQ for funding this research.

So I am a general internist and the problem that we face in primary care with quality measurement and improvement is that we want to routinely measure quality of care for dozens of measures in outpatient practice and use this information to improve care. The cost of chart abstraction for this task would be very problematic and administrative or claims data is really too inaccurate. In particular, those data you really need to capture medical and patient reasons for not achieving a quality measure to get accurate performance measurement and I will show examples of that.

So the potential solution, I believe, is electronic health record systems. They have the potential to routinely measure quality with a high accuracy if a few conditions are satisfied. If you can capture the denominator information - so if diagnoses are entered, if the numerator is entered correctly so you can capture when somebody satisfies a measure with medications being prescribed, screening tests or blood pressure - and then also within the electronic health record you can capture exceptions. So diagnoses, medical contraindications, allergies, adverse reactions, lab abnormalities, etcetera.

About five years ago we started working to do routine quality measurement using our electronic health record systems in partnership with the American Medical Association and the Center for Medicare and Medicaid Services. And this is just the very first early feedback that we were giving to physicians on this. I'll just show you here as one example - this for preventive services is Nunavax (ph 0:06:30) and you can see our rate was only a little bit better than 50 percent, so we clearly had a long way to go. For other things, for example, here was aspirin prescribing for patients with lipid-lowering therapy here and we were doing better. But still, many opportunities for improvement.

As part of that grant we looked to see how accurate the electronic quality measurement was. So the automated measurement was compared to a hybrid measurement where we used the electronic measures, but then if someone failed the measure we actually went in and reviewed the physician notes to find out whether the measure truly was not satisfied.

So first if you look at antiplatelet drugs for coronary disease, which is just one example I am showing here, you can see with the automated review, 82 percent of patients satisfied the measure - so not very good at all. But after the physician review of the notes, that actually went up to 96 percent, so 14 percent higher based on the actual review. So really the automated measure in the electronic health record was not very accurate at all and there are two basic reasons. Aspirin was often prescribed but since it was an over-the-counter medication it wasn't recorded on the medication list despite the obvious implications of that for drug interactions and other things. Also there were many people who had adverse reactions and those were not routinely captured in a queryable format as well.

You can see also for lipid-lowering drugs the performance or the comparison between the two was better, so 93 percent for the automated, 97 percent after the physician review. But there was still a gap of four percent. So in other words, if you had a hundred patients, seven on the automated review would appear that they had a problem but really only three of those seven would truly have a problem. So that means clinical decision support tools are inaccurate and if you are going to do outreach and disease management you will be wasting a lot of time with nurses calling patients who don't need it. So this is clearly sub-optimal from our perspective for quality improvement.

Our conclusion from this work is that overall there is good agreement between quality measured by electronic health record data compared with physician notes, but several factors limit the accuracy of electronic health record measures. And many patients, for example, did not actually have the diagnoses that were entered, so heart failure and coronary disease were put in and then later were found out to be incorrect. Medications, as I said, were not always documented. Exclusion codes that were routinely used in some of these queries sometimes were not valid. So there are a whole variety of issues that came up and challenges. But most importantly was that exclusion criteria often were not captured in a queryable format.

So the question is, "Is this good enough? Is this level good enough?" And when we submitted these papers for publication it was very interesting because we were more critical of the findings and the reviewers were saying, “Gosh, this is pretty good.” But we think that for quality improvement it is really not adequate and I’ll show you why. So this is a complicated graph and I will try and take you through this. This shows you the consequences of missing exceptions and not having those in a queryable format and the accuracy of feedback decreases as performance improves. So imagine a situation where you have a hundred patients on a performance measure.

Early on when you are starting your quality improvement you have got 60 of those 100 patients who meet the criteria and satisfy it. There are another ten that are what we call a false failure - it looks like they don’t need the measure but if you dig deeper they actually do. They have an exception. Then there are 30 people who truly have a failure or what we sometimes call a quality deficit. So what that means is, is the alert - if you were going to get an alert on 40 patients who didn’t need it, the alert would be correct 75 percent of the time. And that is pretty good. We think that that is okay. But now if you go to a situation where now performance has improved up to 85 percent, if you still have those ten false failures and you are trying to just find those five out of 100 to identify them and use your clinical decision support then your alerts are not going to be accurate. For example, in this case you would have 15 people who didn’t meet the measure and only five of them would truly need to have attention paid to them to try and rectify that.

So the alert, for example, if you had a clinical decision support tool, would only be correct 33 percent of the time. That really means that doctors will then start to ignore not only that alert, but they may well ignore other alerts over time if they are not highly accurate.

The implications of this for quality improvement is as quality of care improves, point of care alerts for individual patients are usually incorrect, physicians ignore the alert - if you have doing population-based disease management - lists of patients who need outreach are usually incorrect and outreach is expensive, inefficient and impossible to sustain. This reference here is an actual example where we tried to do that and learned that lesson the hard way.

So the electronic health record can improve measurement by letting physicians document reasons why a patient is not getting an indicated medication or service and actually improve the accuracy of the measurement and that, in turn, improves the accuracy of clinical decision support and other quality improvement tools. So the doctors can document if there is a medical reason - contraindication, etc. - if there is a patient reason, if the patient refuses or can’t afford a medication. One example that comes up for us fairly frequently is that it is a system reason. The most one for us is influenza vaccination. We can’t get it because we don’t have the supply in yet or we don’t - we are out of the supply and we can’t get more for the remainder of the year.

We think that by improving this point of care capture of information this accurate measurement can create a virtuous cycle for quality improvement. So if you are thinking about your clinical decision support, your reminders and these other timesaving tools, you record the exceptions and external data if tests were done elsewhere. That improves the accuracy of the clinical decision support. It also improves the performance measurement and feedback which, in turn, raises clinicians’ expectations, provides more accountability and provides motivation to use the decision support. It actually becomes not only a way to improve care, but it becomes very efficient was well because you only see things that are really important and you know if you are not getting an alert, that something has actually been done already.

So this is all the lead up to the UPQUAL project which was really designed to try and test whether creating this virtuous cycle could lead to actual quality improvement. So our goal was to implement a multi-component quality improvement intervention and the aim, as Angela said, was to achieve ultra-high level of performance through more accurate performance management. We weren't interested in small, statistically significant changes. We really wanted to take our performance to very high levels and then use quality measurement systems to really drive focused quality improvement.

The components for UPQUAL was first our audit and feedback to physicians. I already showed you that graphical report and physicians continued to get that as they had been getting in the past. And then we have these point of care alerts that I will show you for quality measures which are not satisfied and they allow very easy review and ordering and it allows documentation of medical and patient reasons for not ordering. These medical and patient reasons were sent out to a manager and member of a quality committee for review and I will talk a little bit about that. One of the interventions that I think was the most important is we provided monthly feedback to our physicians on individual patients who were not receiving essential medications. They didn't just say, "You are at 94 percent"; it said, here is your actual names and medical record numbers for five patients with coronary disease who are not on statin therapy, for example. And that had a big impact on the physicians. So I won't go through this in detail, but we had a very ambitious goal of trying to improve care for all of the measures shown, multiple measures for coronary heart disease, heart failure, hypertension control, diabetes and multiple preventive care measures. Again, we were trying to improve multiple things at the same time which is just a reality of primary care. We need to be able to focus on a large number of things.

I will show you some examples of how the best practice alerts work and then I will show you some results. First I will just note that portions of the screenshots are hidden Epic's request. We are on Epic, which many of you are probably familiar with and they had asked that we not show all of the formats here. That is why some of the things are blacked out in blue. First point is our alerts are passive. We have no pop-ups with the exception of one for med-med (ph 0:16:45) interactions. But for these quality measures, everything is passive. The physician - if you see a yellow light you know that there is something outstanding. It doesn't mean that you have to do it today but it will stay there until you do something about it.

If the clinician clicks on the best practice alert it opens up all of the different things that are out of adherence. All reminders are through these best practice alerts, including health maintenance. So it is sort of one-stop shopping which is very nice for the physicians. The reminders have these built-in jumps to allow physicians to review key data. So for example for the pap smear alert, there is this jump to health maintenance and they can go and look at past results and look at what the proper interval is, etc. So we talk about this as a "hub and spoke design" that when you get your alerts you will go wherever you need to go or you can go wherever you need to go to get the information that you need to understand how to respond to that.

So in this case you jump and you see that the Pap smear is set to a yearly test and the patient says that they had it done elsewhere. You can just enter in that it was completed elsewhere and then you are done and you enter that in. It really takes about five or six seconds. For this alert here, it says consider beta blocker for heart failure with left ventricular systolic dysfunction. You can’t believe that you forgot something that important so you can just jump immediately and look at the medication history and say, “Oh yes, that’s right. There was a problem in the past. The patient was on Toprol and I stopped it because of an adverse effect” and you put that in and you won’t see that alert.

So imagine a physician sees a patient who actually needs testing or treatment. In other words, there is a true deficit. If you go through this - and I’ll try and go through this as you would feel in real time here. You see the best practice alert. Everything is pre-checked for you. You can un-check something that you don’t want to do. You click “accept”. It has this linked order set that is pretty standard on most programs - but we really tried to pare this down. We wanted to make it simply. No need to read. Everything is pre-checked. So if you are ordering these things, you’ve done it ten times before, you don’t even need to read and find out what you want to check. It is all done. You click “accept”. It puts in the orders. It puts in the diagnosis codes. It puts in the linkages and you are done.

It really makes it very easy to keep up with even multiple things at the same time. If the physician sees a patient who can’t afford or refuses recommended services - in this case the alert is “Consider anti-platelet drug for coronary disease”. There is a tab - you click “Not Done-Patient Reason, Cost.” There is a text box you open up and you can put “intolerant to aspirin, cannot afford Clopidogrel.” And entering that exception suppresses the alert for one year just because things do change. Even if patients refuse, they will sometimes reconsider later on. But that would come up after a year.

Once that information is entered we could do outreach to patients with documented patient exceptions, particularly for refusal of preventive services. So the first year on each week, the care manager received a list of patients who refused a recommended test. They were sent informational materials and called and 6.1 percent of people subsequently completed the preventive services that they refused but when we did an analysis comparing to a different time period we found that that rate was exactly the same as what we had during a comparable time period when we were not doing that outreach. So our conclusion was that the doctors do a good job talking with people about these preventive services and if somebody refuses, then additional outreach and information is not worth it. So we have stopped doing that outreach.

Imagine a physician sees a patient who he or she thinks has a contraindication to a medication. In this case, consider beta-blocker for heart failure with left ventricular systolic dysfunction. Put “not done”, medical reason, type in “symptomatic bradycardia” and again, entering the exception suppresses the alert for one year. In an ideal world we would be able to say, “Tell me again in a year or tell me again in three years or tell me again never.” But that is not something that we are able to construct for this.

One of the things we were concerned about is if we are going to have physicians routinely entering in medical contraindications. Are they valid? We looked at 614 exceptions that were entered. Ninety-four percent were medically appropriate. Three percent were considered inappropriate and three percent were uncertain. All of these cases of the inappropriate exceptions were then discussed at a faculty meeting. For example, one was “aspirin is contraindicated if there is a hemorrhagic stroke or diabetic retinopathy which is not true. All the ophthalmologists say your eyes don’t do you any good if you die of a heart attack. So that was clearly brought up and discussed and I have to say, this was one of the absolute favorite CME activities that we had because all of these were really difficult, challenging cases. We weren’t going over basic things and as a result of that real positive feedback. Now for any new physician who joins the practice they actually have to review these cases to make sure that they don’t have a similar misconception. So this is a good example, I think, of creating a learning health system where we are actually hopefully continuing to advance our care.

It is also very irritating if you get these alerts in a patient who is terminally ill so there is a way to go into the record and just click, “Stop All Reminders-Medical Reason.” And we have validated the accuracy of that as well when physicians enter that.

Our biggest problem now I say, is improving quality for the unseen patients. If somebody comes in we know they are going to get good care, but we’ve got a lot of people were supposed to be coming in and they didn’t come in. So this is what we hoped with these essential medication lists - they will identify patients with diagnoses on the problem list, past medical history, and identify those without a medication on their active list and they don’t have an exception and then we give this list to physicians. We still do this monthly. UPQUAL ended about a year ago and now this is just business as usual. We asked physicians to review the charts and either document the exception or contact the patient. These are really the most important medications. So here is just an example of what we all get and it is interesting. I continue to have people - new patients or something will come up or somebody where they stopped Warfarin and they had been on before and all of a sudden the alert will come up to consider anti-platelet drugs. So it is really a great way to take care at the highest level possible.

So did this do any good? And I’ll go through this fairly quickly. The paper was published in “Medical Care” a few months ago and the reference is in the slide. This is for coronary disease measures and all of these studies were done using time series modeling. Because we are measuring with the electronic health record, we can look at this period - the pre-intervention period. This line here is when UPQUAL started, so you can see for anti-platelet drug it was very flat. We started UPQUAL and that went steadily up. Pretty flat as well for lipid drug and that has gone steadily up. So if I show you the slopes of these lines you can see that we clearly saw an improvement when it was pretty flat before. Here is one that was slightly increasing for ACE and ARB and increased slightly but less dramatically.

Heart failure measures also improved more rapidly and, again, I won’t go through all of these except to point out this anti-coagulation in atrial fibrillation had this very dramatic increase after the start of the intervention. And this was virtually all due to documentation of exceptions. People weren’t starting Warfarin, it was really just the documentation of the exceptions but that really, again, allows you to identify those remaining people who really do need Warfarin.

Diabetes measures improved more rapidly but processes more than outcomes. So for example, here the red line is screen for nephropathy - that improved. Aspirin prescribing, again, improved very dramatically and this was not just documentation. About half of these at least were clearly new initiation of aspirin when it had not been before. But for other things like this hemoglobin A1C

Less Than 8, the proportion stayed pretty darn flat through the study period which was not expected. So sort of a mix in terms of what was improving and what was not.

Beta blocker for patients with myocardial infarction - that improved at just about the same rate so there was continuing improvement but UPQUAL didn't appear to boost that rate of improvement at all. The prevention measures improved also at pretty much the same rate. Cervical cancer gradually increased. Pneumococcal vaccination was already skyrocketing. If you remember back - we were at about 50 percent when we started giving docs feedback so that was going up at a steep rate and has continued to go up to about 93 percent now. And colon cancer also showing a slow improvement. So less impressive are results for the prevention measures. I think partly that is just because the previous alerts and feedback was working so well.

I'm going to skip over this for the sake of time and I'll skip over these two.

So in summary, for the first year of the UPQUAL intervention 14 of 16 measures improved significantly. Nine measures improved faster than over the preceding year. Four others improved at the same rate compared to the preceding year. One improved, but at a somewhat slower rate, and one did not improve. One actually decreased, which was mammography, which was really related to internal problems with long delays in getting mammograms so we don't believe that that was related to the intervention itself.

The key lessons that we have learned from UPQUAL is first, health information technology is really just a tool to execute your quality improvement strategy. It is not a strategy in and of itself. But if HIT is used to support a comprehensive quality improvement strategy like UPQUAL, care can be significantly improved, but these clinical decision support and other QI tools must be seen by physicians really as their own personal QI tools. And that is what has happened really in the last couple of years, that the docs said this is the way they do business. Last summer it thrilled me when they were asking for more alerts and more measures because they say if it is not included in this panel then they tend to ignore it. So that is sort of the final hurdle so that this doesn't feel to clinicians like this is somebody looking over their shoulder. It is really their tool.

So I will close with that and with this, my favorite cartoon that says, "You know, you can do this just as easily online." And I will turn this over now to Andrew Hamilton.

ANGELA LAVANDEROS: Hi, this is Angela Lavanderos. Now we are going to hear from Andrew Hamilton. Mr. Hamilton will present findings from the AHRQ-funded project entitled "A Partnership for Electronic Health Records Use and Quality of Care." He will focus on quality indicators identified by the Institute of Medicine for areas of improvement through rigorous quantitative methodology.

ANDREW HAMILTON: Great. Thank you. And like Dr. Baker, I want to thank AHRQ for the opportunity to present today, but also for funding our project. So I thought I would provide a little background in terms of the work at this particular project for those that are not familiar with health center controlled networks. The Alliance of Chicago is a network of community health services that really developed a joint venture to come together mostly focused on efforts to improve quality and safety. The joint venture led us toward the implementation of an electronic health records system in that network.

In addition to that, the Alliance of Chicago has partnered with the Institute for Nursing Centers, which is a similar organization that was founded by funding from Kellogg to bring together nurse-managed health centers for the purposes and very similar efforts around improving quality and safety as well as looking at issues regarding sustainability of services as nurse-managed centers are rapidly becoming a new and growing source of primary care in the United States. The Alliance and the Institute for Nursing Centers partnered to develop this AHRQ project which was really to examine how a partnership approach could serve as a purpose to promote full use of electronic health record systems with the end goal of improving quality and safety of care amongst these safety net practices. Our specific efforts were really looking at linking clinician use of EHR against how clinicians were scoring, as it were, on quality of care measures relevant to preventive care, chronic disease management, and medication safety. Then secondly, looking at how the partnership as an organization process helped drive implementation and how our work really has, of course, predated the development of the Office of the National Coordinator of Regional Extension Centers but yet the model of the partnership really does look very similar to regional extension centers. So our work is also hoping to provide more detail and evidence around how those partnerships can lead towards quality improvement.

What I will say about today's presentation - because the project is fairly large, what I have decided to do is focus specifically on a subset of data that we have collected regarding the medication safety component of electronic health record systems. As we all know, one of the goals of health information technology and improving safety is targeted specifically at the medication safety features in electronic health record systems. Initially our project had just a few measures where we were looking at basically clinician satisfaction with medication safety prompts as well as looking at the documentation rates for ensuring that allergies are being documented appropriately. What we found in this preliminary data was that, in fact, clinician satisfaction was relatively low as it related to medication safety features. So we decided to do some further data analysis and collection in this area. That is what I am going to share with you today.

Just a little bit about the methods and the setting - so essentially there were three nurse-managed health centers that were selected as the intervention sites. I have listed them here, so Applied Health Services, which is a federally-qualified health center serving homeless populations in the Tenderloin neighborhood of San Francisco; the Campus Health Center located on campus at Detroit Michigan at Wayne State University, which is providing primary care largely to international students and now a growing number of extra local population in the City of Detroit; and then Arizona State University nurse-managed health centers which are providing services through three cities in Phoenix, Scottsdale and Tempe largely uninsured primary care adult services.

The federally-qualified health centers that participated were Howard Brown Health Center which is an urban, HIV, gay, lesbian, bi-sexual, transgender health clinic in the City of Chicago; Erie Family Health Center, another urban clinic serving mostly Hispanic and recent Mexican and Puerto Rican immigrants; and then Heartland Health Outreach which is a health center for the homeless located in Chicago. These three sites served as our comparison groups. Prior to the launch of this particular study, these three sites had already been live on the electronic health records system and the three previous sites were in the early phase of pre-implementation.

The data that I am going to walk through today is sort of in three buckets. So again, we look at quantitative data regarding system use so our attempt was to measure how frequently, in this case, the drug-to-drug and drug-to-allergy interaction checking feature was being used within the electronic health record system. We also measured user satisfaction with that tool and then specifically looked at the percent of patients known to have allergies documented just to obviously have a control in terms of whether or not the drug-to-drug and drug-to-allergy checker could work. We found very high rates of allergy documentation - actually 98 percent and above of all patients that ever had a visit through the measurement period had allergies documented so it ended up not being an overly interesting variable from the perspective of this particular study.

We also analyzed from our key informant interviews some qualitative data that I will share with you. And then lastly we observed both enterprise settings as well as user-specific settings for the drug-to-drug interaction checking. So for those that are familiar with commercial electronic medical record systems, in the typical scenario you can have the drug-to-drug interaction feature active or not active and then once it is active, in many systems, the user themselves can control how sensitive that interaction checking tool is and how accurate it is as it relates to data that is available.

So from the quantitative data, essentially what we looked at is we queried the database for known drug-to-drug interaction pairs. Essentially, the way that was defined is for any known drug-to-drug pair that had overlapping start and stop periods they were considered a possible match. And we also specifically - for the control sites we limited our review to only medications that had end dates in 2008 or greater. So if a medication had previously been entered in 2005 or 6 and then stopped by 2008 we didn’t include that in the comparison because none of the intervention sites were live at that period of time.

What we used as the baseline information was from a CMS tool that was published that listed common drug-to-drug interactions so this tool allowed us to identify those that are known drug-to-drug interaction pairs and then we also, during the preload period at all three of the intervention sites - had the preload staff indicate any time they were preloading a patient’s medication list and when they were doing that preload, if they received a drug-to-drug interaction prompt. That was a way to, by proxy, understand the rates of drug-to-drug interactions pre-implementation.

I am going to share some of the results of that data in just a moment. As I mentioned, we did not set out to do that type of query as part of the initial scope of the AHRQ project. What really led us there was some data related to user satisfaction. I mentioned that we measured user satisfaction. We are using a standardized tool that we have actually published in previous literature that we have done psychometric (ph 0:38:14) evaluation and analysis of the tool. This tool is being implemented - during the implementation period, at the six to twelve month post-go live and then at two years post-implementation. What you can see on this data, there are sites that have not hit the two-year post-go live yet. They will be within the next two months or so, so we will be able to finish this data analysis.

The trends overall, of this data, is that the post-implementation evaluation so that period at two years out typically rebounds or, in fact, is higher in some cases than the baseline data which is measured right at go live. So not overly surprising that users as they become more comfortable with the electronic health record system would likely become more satisfied. What we did find, however, is that in specific areas the tool is a multiple number of variables. One of those areas that we focused on was around the drug-to-drug interaction features. So users have reported both in the qualitative data as well as our key informant interviews, very high expectations around the use of the drug-to-drug interaction feature. While they rate it as highly intuitive and easy to use, unfortunately the usefulness of that tool was rated very low. In fact, it is one of the lowest indicators on all of our user surveys and that is what prompted us to really understand what was happening and learn more about why this particular feature in the electronic health record system, which is really a very important feature for safety and quality purposes, was not perhaps performing the way that we had hoped. So we added some questions to our key informant interviews and what we found is that in many, many cases, the drug-to-drug interaction alerts are, number one, very infrequent, which I can validate with our quantitative data that I will show in just a moment. But overwhelmingly users are not being frequently alerted that medications they are using are known drug-to-drug or drug-to-allergy indicators. However, what does occur is when the prompt is alerted to the user, the clinical relevancy of that prompt is often in question.

Some examples that we found by our key informant data really underscored that in some instances, of course, there is clinical relevancy to using medications in which there is a known drug-to-drug interaction. So with that were the examples that were provided sort of - antibiotics was Warfarin or psychotropic medications with other common interactions. Birth control pills which are obviously a problem in our campus health center sites with other medications. So what the general theme from the data is that the drug-to-drug interaction alerts do not in any way distinguish between the serious or very significant drug-to-drug alerts against those common drug-to-drug interactions. So what we found is users wanting to have the feature of the system be more specific in how it alerts the user around these particular pairs.

Lastly, what was also illuminated through these interviews is that users are not able to, in our current commercial electronic health record system, are not able to document reasons for overriding the alert…in a standardized way. Of course, they can document in the note the reason for overriding the alert. However, from a data collection and sort of research perspective, we are not able to analyze that data in a structured format so that becomes difficult in terms of being able to do anything more significant with the data.

Lastly, one other theme that appeared is the need to be able to obtain more information about alerts that may be obscure or not well known so that, for example, the users would love to be able to navigate very easily to the pharmacological information about that known alert.

So just in terms of the data itself - what we found when we actually did run the query of, again for the two years of data that we analyzed, there were only 645 drug-to-drug interaction pairs documented across all sites in the study. That number rests against the 64,000 unduplicated patients that were seen by these six clinics across the period of time of that study. So this is a relatively low rate of known drug-to-drug interaction pairs. However, we assumed from the qualitative data that many of those known pairs actually were often a temporary clinical necessity and we wanted to be able to quantify what we found in terms of those known clinical necessity. Secondly we wanted to know were there actual serious safety considerations that were missing or that the health information technology system is not helping us be able to analyze. So we did some further analysis on the data with the question in mind, “Is there a real medication safety concern or is this an artifact?” So it is very similar to what Dr. Baker presented around the need for precision of both the clinical decision support tool, but also the need to be able to find for those outliers, or in this case, those safety concerns, the ability to go all the way and narrow down to those that really are important for us to actually do something about.

So the first thing we noted is that about 88 percent of those in the sample - of that 645 pairs - 88 percent had a missing end date on one or both drugs. What that means is that we can’t actually be certain that the patient is actively taking the medication that is still active on the medication list. Perhaps with the more robust features of a full e-prescribing system or a medication history or prescription refill communication with the pharmacy we would be able to do this. But in the context of the data we had it was very, very challenging to know that those 645 actually represented a quality issue.

From that - of that list then of the 565 we narrowed it down further and found that about 53 percent of those had one drug with no end date and the other drug in the pair had a start date before 2008. In other words, we could assume that the start dates were so far apart that even though the end dates appeared that they were overlapping, it is potentially that we could assume that, in fact, those 342 of that 645 weren’t relevant or were probably not safety issues. But again we can’t say that as 100 percent accurate. And what we also found in addition to that is about 214 actually had start dates that were within one month of each other. So those are the ones that were maybe highly suspect as representing true quality issues. When we further analyzed those we found about 19 percent, so only 120 of the total 654 that we looked at actually represented what we considered likely safety issues or safety concerns. So the point here is two-fold. One it is important feedback for developers and commercial software design to understand and appreciate the nuances of the precision that we need for drug-to-drug interaction and secondly it points out the potential system use issues around - in this case - instructing users and training them around the importance of entering start and stop dates within an active medication list. As I am sure we are all aware of the components of meaningful use and specifically measures around the up to date medication list. We are very concerned that these technologies become enhanced to be able to better manage this data.

Ultimately of this 19 percent that we then provided feedback directly to each clinician in the study - we gave them a list of patients who were in this group of 120. We really want our clinicians to have tools to be able to respond and address issues that are potential safety concerns. Obviously if the decision support feature were more relevant and less ambiguous it would help in terms of that and that work.

The high-level finding here, again, the decision support feature is not always reliable and does not eliminate potentially harmful combinations because it provides sometimes inaccurate or clinically necessary reasons why patients are on drug-to-drug interaction pairs. So that was the thesis or finding. The last thing that we’ll say was a major limitation to our work was that the drug-to-drug interaction feature itself - so the prompt that the clinicians see - does not actually store a marker, if you will, in the database. So we are not exactly sure how often clinicians are receiving these prompts. Although, again, in our qualitative interviews it was clear that it was very often. We didn't have a direct way to measure other than saying how many medications actually resulted in a drug-to-drug interaction. This would be another hope for future work, to be able to really have more precision around capturing when that alert, that decision support prompt actually does present itself, that there is some way to go back retrospectively and analyze that just to understand how this particular tool is being used.

We also wanted to have one more data point around looking at safety practices amongst the participating health centers so part of our initial research design involved the deploying the Physician Practice Safety Assessment or the PPPSA which is a tool that was developed by previously funded AHRQ work. So we implemented this as a self-report safety assessment tool at the pre- and then post-go live periods and - our attempt was to compare overall safety practices pre- and post-go live in the context of using health information technology. We will be publishing work in this realm because we found some very exciting things relevant to both quality improvement work as well as safety practices and how technology is definitely a force to support your efforts there.

We are also going to look specifically at the qualitative data from this medication safety feature against specific indicators on the PPPSA to understand to which extent how IT may have helped drug prescribing practices within these vulnerable settings. So with that I will end my presentation. Thank you.

ANGELA LAVANDEROS: Thank you very much. So now we will hear from our final presenter. Dr. Wiener will explore the findings and implications of his AHRQ grant to use EHR data to improve quality measurement. He focused on diabetes through this project. Dr. Wiener will describe the shortcomings of the existing HBA1C threshold-based quality measure and propose an alternative measure that credits providers who are performing better than expected.

DR. MARK WIENER: Thank you very much. I, too, am grateful for the opportunity to speak with everyone today. While the two prior talks were discussing technical and logistical mechanisms for improving quality of care, particularly with respect to process measures and avoiding drug interactions, I am going to discuss today the ways you can use information from electronic health records to measure and compare the quality of care based on outcomes and some issues to consider when developing and applying these quality measures.

The motivation underlying this work arises from a general sense that, prior to the availability of electronic health data, the data that was being incorporated into quality assessment scales was chosen on the basis of its accessibility rather than it being optimally suited for the quality measure. A classic example of this is with diabetes when only the most recent clinical parameters are considered in the assessment without any accounting for prior values and the trajectory of the parameter over time. While quality measures assessed in this manner are clearly related to quality of care, the fundamental flaw relates to the fact that measures, by their nature, are cross-sectional, whereas diabetes management inherently is a longitudinal process. But with electronic health records it becomes feasible to capture the longitudinal change in clinical status of patients and capture and adjust for many more relevant variables than just the outcome of interest.

Our study sought to incorporate electronic health data into models of quality to see if the new models perform better than old models. However, we gave a lot of thought to the notion of what performing better actually means. While there are some objective measures of what makes a good doctor, the notion of quality care, as you can imagine, has a lot of subjectivity to it. Indeed, what makes a good doctor? And who is the best judge of a good doctor? What are the relevant metrics of a good doctor? How to you compare the quality of care of two doctors, especially when the environments in which they practice can be very different? How should the characteristics of patient served by a doctor be incorporated into the assessment of quality of care? Lastly, is the best doctor truly the same for all people?

In other words, when you create a ranking of physicians why wouldn't everyone want to go to the so-called best doctor? So any discussion of quality must make mention of Donabedian who laid the groundwork for the framing of the notion of quality of care. He defined the four axes of quality - structural measures, which look at the appropriate credentialing of staff and things like board certification. That is pretty fundamental but nonetheless very important. Next are satisfaction measures which are the patients' perceptions of the relative benefits of treatment and the quality and quantity of life, balanced by the difficulty of undergoing the necessary treatments. This is an important aspect of quality of care because sometimes the more aggressive treatments that earn better quality scores can actually make the patient feel worse and feel less satisfied with their care.

Thirdly, process measures assess the degree of adherence to standards of practice and lastly are outcomes measures which evaluate clinical endpoints such as functional status, mortality, or hospitalizations as a result of treatment.

In the interest of time I am going to focus today on what we covered through our funding which was outcomes measures. The pros of outcomes measures is that they reward the truly tangible benefits of the care practice. But on the flip side, any real change in outcome is going to often take years to develop and it is difficult to detect statistically meaningful differences. Furthermore, many outcomes are highly dependent on patient behaviors and other conditions that are beyond the control of providers. While we can probably debate what are good outcomes measures for diabetes and whether or not death and hospitalization should be in the forefront, I do accept the general notion that A1C, LDL and blood pressure are reasonable parameters to measure for diabetes. But we also recognize that, strictly speaking, these are not true outcomes and are better considered as intermediate outcomes.

So the basic approach to quality measurement for diabetes says that you are a good doctor if you are doing all the correct process things such as foot exams, eye exams, urinary micro (inaudible) testing. But you are also measuring lab parameters and achieving good results. You are a good doctor if a high proportion of your patients with diabetes have a most recent A1C value less than seven, an LDL less than 100 and a blood pressure of less than 130/80. Of course, these are the clinical targets but the quality measures tend to build in some leeway when the thresholds for quality are somewhat higher than the clinical targets. But the net implication is the same. The more patients you have with lower values, the better you are. A corollary of this is that you are an improving doctor if your score this year is better than your score last year. But it doesn't take much to think about many ways this can happen without any real change in the true quality of care.

Even with these relatively straight-forward definitions of quality, there are a variety of issues around organizing data that clinicians would want to have recognized in assessing the quality of care. Fundamentally, who should could as having diabetes? Physicians may say, "Well my patients can't be treated to targets because they keep having hypoglycemic episodes." Or they may not be succeeding in achieving the targets despite being on a lot of anti-diabetic medicines. Or the measure is not accounting for the fact that the patients are sicker by some measure, or it's not my fault the patients are not succeeding because they are just not compliant with their medicines. Or well sure, the recent A1C value was poor, but the last A1C value was excellent. Lastly, they may complain that they are just really busy and they need more team support with addressing all the issues underlying the panel's poor control, so it's not necessarily a provider issue, but a team issue.

Other structural issues that you need to account for in assessing quality is do I have a large enough panel to reliably assess the quality? Have I been responsible for the patient long enough to have an impact? A third one, are these patients really mine, which turned out to be a very huge issue in the correct attribution of quality. And lastly, are there factors of success that are really more the patient's responsibility than my own as the provider? And how can you measure and account for that?

Regarding the first question of who should count as having diabetes and should this definition match an epidemiological definition that really strives to count everyone who could possibly have diabetes in the country. Well the reflex answer sounds like it should be, "Well yes. Everyone with diabetes should count in the quality measure." But you can see the implications that if you label some patients who are barely diabetic as having diabetes you may improve your quality in terms of A1C because these barely diabetic patients, sort of by definition, have good A1Cs, but then the implications of labeling them as having diabetes would also require that these same patients meet the stricter LDL and blood pressure goals associated with diabetes and these barely diabetic patients may not meet those criteria. So the impact on an overall score may be not as good as anticipated. Also if I know my worse-controlled patients are going to count I might, maliciously or otherwise, send them away to others and that could improve my quality score.

So we began some rudimentary analyses to explore some of these notions and their impact in greater detail. One of the most basic analyses we did was simply looking at the case definition of diabetes as a function of the number of diabetes diagnoses they accumulated. So this table shows there is a gradually increasing trend in average A1C as the number of accumulated diabetes diagnoses increases. Perhaps this is not surprising since we know that diabetes does get harder to control the longer one has the diagnosis and more diagnoses are simply a marker for having the disease longer. But not necessarily if you pack a lot of visits in a short period of time. But in any case, these findings do suggest that unless we are correcting a quality measure for the number of diabetes diagnoses we may unfairly penalize doctors with many long-standing diabetics. Of course, patients with a few diagnoses may have had long-standing diabetes managed elsewhere recorded in another information system but the data we are presenting average across all patients. Unfortunately we don't have a good sense of when diabetes was first diagnosed in our patients, at least through the EHR.

We went on to look at other parameters that have been promoted as being significant in terms of the epidemiological definition of diabetes. We looked at the impact of diabetes medication use among patients who otherwise have an established diagnosis of diabetes based on having a diagnosis recorded at least twice. And perhaps not surprisingly, among patients who were never on hyperglycemic medicines, the A1C average was 6.23, whereas patients who were on at least one diabetes medicine had an average A1C of 7.36. That is a huge difference. Other authors have suggested looking for patients with inpatient diagnoses of diabetes. However, interestingly, if you look at patients whose only reference to diabetes was an inpatient, the average A1C of those patients was 6.6. This contrasts with patients whose diabetes diagnosis was an outpatient, where the average was 7.18.

Another approach could be defining on the basis of an elevated A1C. But the problem here is that stacks the deck against the ability to have good control since the inclusion into the measure was conditioned on the A1C. Therefore, although it feels good to have a very broad definition of diabetes, there is a lot of inherent heterogeneity to that kind of definition and therefore it does make sense to perhaps narrow the field a little bit. We ended up including those with at least two outpatient diagnoses for diabetes.

As I mentioned earlier, the key problem with the current outcome measure for diabetes is that it only looks at a point in time assessment of A1C without accounting for change from prior levels or other clinical factors. There is no accounting for patient-level characteristics. But you need to be careful when you do that because you need to avoid two things. One is gaming of the systems, such as recognizing that depression maybe be a reason for the difficulty in controlling the diabetes. And then if you account for this either by simply not counting these patients or allowing them a different A1C threshold for quality the problem is that it is too easy to label patients as having depression simply to get a bye on the diabetes measure. And not everyone with depression is equally as effected by it in a manner that might affect the diabetes. Furthermore and perhaps even more importantly, is the need to avoid the impression of a double standard. If patients with depression are found to have systematically worse control and you adjust for it, then providers of patients with depression can seem to provide high quality of care while essentially allowing patients with depression to have worse control. While it may be okay to allow selected patients to have worse control for legitimate reasons, it is not a good idea to make such a broad generalization with large groups.

Another absent feature of current quality measures is that there is no accounting for provider effort. Provider effort could be assessed as the process measures, but in this case I’m talking about medication prescribing and intensification. But still we need to avoid gaming of the system that would allow disingenuous medication prescribing just to look good. Sometimes a doctor can write a prescription but not really support its use. We really need to avoid that.

Lastly, we need to be mindful of the potential unintended consequences of sub-optimal quality measures. For example if higher socio-economic status predicts better control and provider of easy diabetic patients in the rich suburbs receive the pay for performance bonuses to the exclusions of providers in the so-called hard diabetic patients in the urban poor community. The impact of that is instead of working to improve the care of urban poor patients, these doctors may flee to the suburbs. Of course in reality, you know there are reasons this doesn’t happen but it certain is a theoretical issue. Furthermore the apparently high ranking providers may attract more difficult patients for which the so-called best provider has little experience.

Other issues we tried to think about before we even proceeded to conduct any analyses is where and how to set the threshold for quality. In considering this issue we asked ourselves, “What is the goal of the quality measure? Are we trying to recognize or remediate poor performing providers? Are we trying to reward good providers? Are there clinically meaningful differences between a highly ranked and a lower ranked provider?” Related to this is the panel size issue. Can a good or poor measure in one patient in a panel skew the overall quality measure? And lastly, the criteria you set for quality should be clinically important, definitely, but it should have good discriminatory characteristics. Therefore if everyone can achieve the goal, by its nature it should carry less weight even if we agree that it is an important part of diabetes management.

So in thinking about these issues we came up with a framework for a novel solution. Rather than ranking providers based on the proportion of the panel with good control, we sought to create a level of expectation for the clinical parameter of values and then ranked providers on the degree to which their patients are doing better than expected. We recognize that this method means that some patients having certain characteristics would have a lower expectation of control. However unlike the condition I alluded to before, this is not a double standard because we would ensure that maintenance of the status quo is not rewarded. You must improve control beyond expectation to receive quality bonuses.

We also envision that providers of so-called easy patients with good control who are expected to have good control wouldn’t automatically be labeled as poor doctors, but nor should these folks be considered the best doctors. To receive the best label, they really need to take on some riskier patients and improve the control in these patients beyond expectations.

So with this framework we began our analysis with patient selection. We included patients with at least two diabetes diagnoses from eleven primary care practices that were already on the electronic health record. At the time we started this study we looked at visits between January 1, 206 and December 31, 2007. To address the issue of having enough time to address an elevated A1C, we required patients to have a most recent A1C drawn between December 2006 and November 2007 and a prior visit within one year of the current A1C. We assigned patients to providers if they were seen between 90 days and one and a half years prior to the current A1C. So basically we were giving doctors at least three months to recognize the poor A1C in the past and act to improve it. We also limited the sample to patients and providers having at least ten patients in the panel. Although this was somewhat arbitrary it was meant to avoid the problem with panel sizes that were too small where one patient could drastically alter the percent of patients with good control. The net result as a group of 4,800 patients seen among 92 different providers and this did include residents.

The characteristics of the patients shown here - they were consistent with our overall practice patterns. There were more females than males and the number of blacks exceeded 50 percent. The average age had some variability but it was generally in the low to mid 60s.

The next slide here shows - it still has our multiple significant digits. I apologize for that. The next slide shows the average A1C, systolic blood pressure, and LDL levels of the set of patients. The overall averages actually were quite good, but by themselves the averages masked the fact that a good number of patients did have poor control. One of the things we wanted to get out of the way quickly was this notion that depression can somehow effect A1C control. Unlike what has been reported elsewhere, in our population, the overall average A1C was not markedly different among patients with and without depression. There were racial differences, but within a race the presence of depression was not associated with a markedly different average A1C.

Our next analysis is a very basic one which arises out of a common reaction when providers see rankings based on A1C less than seven, which is, “Well gosh, so many of my patients have values just over that threshold and if only you counted them as having good control I wouldn’t look so bad.” So this chart compares the rankings of providers when you use an A1C less than eight versus when you use an A1C of less than seven. The better ranked providers have the lower numbers and are in the lower left portion of the slide. Even without a quantitative statistical correlation you can see that the rankings are in fact correlated but some provider rankings will change markedly depending on the threshold you choose. In particular the providers whose panels fell along the X axis, those guys do not care about statistical correlation. Particularly this guy whose ranking dropped substantially with the change in threshold. Seeing a graph of ranking creates an artificially large distribution of points that graphically seems to imply that the high ranking providers are much, much better than the lower ranking providers. This graph now shows the rankings not as rankings, but as actual proportions of patients in a panel below the threshold. While the best and worst providers are clearly separate, you do get a better sense of the clustering of points when the data is displayed in this way rather than with the rankings.

To account for the variability of an individual clinical parameter, some authors have suggested creating composite measures of A1C, blood pressure, and LDL control. However for this to really work, the directionality and control of all three parameters should be somewhat consistent. Unfortunately our data did not demonstrate this consistency. So this graph shows a comparison of ranking assessed by having A1C of less than seven against having the blood pressure of less than 130/80. As you can see, the correlation of values is much less apparent than it was in the A1C eight versus A1C seven graph.

A similar pattern of no correlation is seen when you compare the ranks of LDL control and A1C. The pink line is not a regression line but it is simply the 45-degree line that should be the point of convergence for all the dots if there were good correlation. Again, no statistical correlation here, but these dots do not follow anywhere close to that line.

If you take A1C out of the picture and you compare blood pressure control to LDL control, as before you can see a tighter clustering that implies overall less difference among providers than the rankings would suggest. A similar clustering appears when you look at the average proportions of patients with good LDL and A1C control.

And this next slide is a similar figure showing the clustering of providers having proportions of patients with good blood pressure and A1C control rather than rankings. There are a few clear outliers, but many more are clustered more closely than, again, the rankings would suggest. So these were simple unadjusted models that looked solely at the blood pressure, LDL or A1C values. So the goals of the grant - we sought to test the - predict the value of other putative independent variables that were available from the electronic health record. And we looked at the usual ones like age, race, and sex. We did median family income based on the U.S. Census which was racestratified within a zip code. But from the electronic health record we are able to look at body weight and other vital signs. We looked at the number of diabetes diagnoses, we looked at the individual comorbid diagnosis categories by mapping their history of ICD9 diagnoses from billing codes into the AHRQ clinical classification system. We also counted the number of comorbid diagnosis categories. We looked at the types and numbers of diabetes medicines ever attempted. We thought that accounting for the number of medicines ever attempted was more meaningful than that number of current medicines because we believe if a provider had recommended a drug that was subsequently stopped, that there was a legitimate clinical reason why it was not continued. We wanted to account for that in our assessment of medication intensity. We gave a lot of thought to the meaning of the data represented in this graph and how to incorporate it into our analysis. What the graph shows is the percent change in hemoglobin A1C as a function of the two-year prior A1C average. Now granted, the number of A1Cs drawn in the last two years is highly variable per patient, but nonetheless the trends here are very interesting. The blue curve is a threshold where the percent change moves the A1C above a value of seven going up. The yellow curve is when you cross the threshold where the percent change moves you above an A1C of nine.

What this shows is that if your baseline A1C was less than seven you had a 96 percent chance of your most recent A1C also being within 20 percent of the prior value. If your baseline A1C value was in the seven to nine range, you had an 84 percent change of remaining within 20 percent of your initial value. Lastly, if your A1C was greater than nine, you had a 65 percent chance of remaining within 20 percent of your baseline with 31 percent of people actually doing more 20 percent better and four percent of people actually getting worse. And this fits within the biological ranges of A1C.

We interpreted this to mean that while having an A1C of less than seven is clinically desirable, sustaining an already low A1C is also relatively easy to do and it was common enough where it was true in the overall well performing providers as well as the providers that, by other measures, looked like they were performing poorly. So we wrestled a great deal with incorporating the prior A1C as an independent variable because of the concern of introducing homogeneity into the model. On the one hand, prior A1C is an appropriate variable if you consider that average to be an integrative parameter that represents the net effect over all the clinical and behavioral issues. On the other hand, it is possible that the patients with poor A1C cluster within panels of poor quality doctors that really should not be rewarded. In the end we chose to accept the parameter because of the near universal finding that A1C was - a good A1C was relatively sustainable across providers.

Other variables included age, pulse, income, use of diabetes drugs. But interestingly, no diagnosis category made the cut of being predictive in models that accounted for potential random selection of variables that may look predictive in one modeling iteration but not in others. So we used our models to calculate an expected A1C that was based on prior A1C, age, pulse, income, markers for the use of various diabetes drugs, and then we looked at the residuals with respect to the A1C values. And then we ranked the providers based on the sum of the residuals, then we compared the rankings of providers based on the proportion of a panel with an A1C less than eight against the new model which looked at the actual versus expected A1C value. As this figure suggests, again there is some correlation but if the providers whose rankings switched a lot - if their ranking got worse they are going to be suspicious of the new measure.

As I indicated earlier, the assessment of better quality measure has some subjective and objective components. So is this new method better or is it just different? We believe the construct of the new quality measure fits better with diabetes and incorporates the longitudinal aspects of diabetes management. It essentially justifies what many authors have suggested in valuing the improvement in hemoglobin A1C even when the A1C does not cross the usual thresholds. It recognizes that while sustaining an A1C of less than seven is clinically important, it is relatively common across providers so this new method values this achievement less. We could debate whether or not that is a good idea. It also incorporates all patients regardless of comorbidities. It does not say we are not going to count certain patients with certain conditions or that all patients with certain conditions are going to be given different thresholds. Still, though there are many unresolved issues with regard to the new method. As we looked at the panels of patients who are consistently ranked high or poorly under both methods and we also looked at people whose rankings changed, we realized that we may be overvaluing large improvements in one or two individuals over modest improvements in larger numbers of patients. Furthermore, confidence intervals around the expected A1C values were quite large which means that most providers, except for the highest and lowest ranked are statistically indistinguishable. We clearly need to have better adjustments for panel sizes where, even though we required a minimum of ten patients there were several providers who were looking very strong when really only three or four patients had an A1C value better than expected. But that represented 30 to 40 percent of their panel whereas doctors with larger panels had a harder time reaching a 30 to 40 percent of doing better than expected. That perhaps seems unfair.

Our ranking system did not address the issue of patients where the provider never ordered a hemoglobin A1C. While you can argue that that could be poor quality of care, in reality this is done in patients who have a pattern of consistently good control or, on the other hand, patients who have consistently poor control and their finger stick in the clinic is in the 300s and they told me they haven’t been taking their medicine for weeks. I usually don’t bother checking the A1C in those individuals. But I do yell at them.

The attribution of the correct provider is very difficult and we tried many business rules to assign patients to provider panels correctly. But a uniform complaint among our providers is that our algorithms that assign patients to providers do not reflect reality. We tried assigning patients on the basis of who would see that patient most often or most recently. However that approach was similarly panned by our providers. As a result of these, we are making a more concerted effort to actively complete the primary care provider field that is available within our electronic health record. The effort was almost hijacked by our billing department who, in discussions with the insurance companies, wanted to take the PCP field and fill it in with the provider that was listed on the patient’s insurance card. However, empiric evidence suggested that the insurance company’s notion of the primary care provider was less accurate than our own. So in a rare but significant step where clinical importance trumped the financial priorities, we think we won the battle where the PCP field could reflect our own providers’ notion of who they are responsible for. One can argue that the insurance company should be taking their primary care provider cues from our electronic health record rather than vice versa. And we feel so strongly about this we would like to see the simple notion of agreement regarding who constitutes a provider panel as an independent quality measure.

Another issue that our providers had was not wanting to be responsible for patients not seen in over a year. This too was an interesting thought experiment because if the patient was last seen over a year ago and had poor diabetes control and that patient hadn't been seen in over a year, doesn't that reflect poor quality of care by itself? Certainly the longer it has been since the patient had been seen, the more likely the patient might have transferred their care elsewhere or may even have died? But aren't these patients the ones who have the highest risks since they may be falling through the cracks of health care? Perhaps they have landed within another clinical practice but given the potentially higher risk of poor outcomes of these patients who aren't even being seen, it is reasonable to consider the active assessment of the status of patient affiliation with a practice as another independent quality measure.

Regardless of these more general quality measures, the implication of this new quality measure are that providers who had succeeded in moving patients from poor control to better control will be ranked highly in a good way. But once success is achieved, these providers rankings will drop if the panel remains constant and the panel is simply sustaining the average A1C at the level of expectations. Therefore, the only way to sustain a high ranking is to continually take on and succeed with new poorly controlled patients. We think such an approach will help fight the perverse incentives that make it less desirable for providers who happen to care for the poorest controlled patients. Again, taking on these poorly controlled patients doesn't automatically earn you a high rank, it merely gives you a better opportunity to achieve that higher rank.

And I guess we can open up the floor to questions and I wanted to acknowledge my co-investigators on this work. So thank you very much.

ANGELA LAVANDEROS: Thank you. Yes at this time we are going to open up the floor to questions. There is a questions tab on your screen that you can use to submit questions. We are coming up on our time but I would be happy to take questions.

We have two questions that are geared towards Dr. Baker's project, actually. Dr. Baker, you mentioned patient participation. So there was some curiosity about patient feedback that you received, if any. Can you speak to that?

DR. DAVID BAKER: I'm not sure that I understand the question about patient feedback. Patient feedback to the outreach - I think we don't really have data other than to say that it obviously was not well accepted. Again, this was done through letters, sending educational material and then actually having the patient education specialist calling people up. The general feeling was that the doctors had talked about this and, you know, I really don't want to hear from you again. So we think it is important to respect a patient's right to refuse these different services if it is an informed refusal. So that is why we think it is important to be able to document this. I think the fact that patients did not appreciate this is evidencing. We should just be documenting and moving on.

ANGELA LAVANDEROS: Okay. The only other question - and folks can send clarifications if they would like. The only other question was could you speak to the affordability and usefulness for something like this at a smaller practice.

DR. MARK WEINER: I think realistically this is something that is going to be up to the electronic health record vendors to design tools. I think the design of these things and the implementation that we had to go through to get this to work is not something that small practices could do. If they are part of a larger organization or, for example, the regional extension centers for health information technology may be able to provide more support. Actual performance of this in terms of the physicians being able to do that and the feedback reports - we are trying to do that online and automate that now so, again, even smaller practices, if they have that level of support, it should be very easily sustainable and very low cost. The problem is the initial build.

ANGELA LAVANDEROS: Okay. Actually have another question from the audience, again for Dr. Baker. How do you handle reporting outcomes for exceptions? Do they get included in the denominator?

DR. DAVID BAKER: So the most important thing - when we actually report it out to doctors they get their total denominator - number eligible - they get the number who satisfy it and the number of exceptions. So we like to have that transparent and if we had our way that is the way all reporting would be. Again, nationally those exceptions are supposed to be removed from the denominators so that is the way we have done it in our papers. But we think it is important to be transparent. It is important to look for outliers. If somebody is reporting a much higher rate of patient refusals or medical contraindications than other providers, then people need to look into that. That has not been the case for us and for much larger projects, but I like the idea of having the exceptions being very transparent and not just pulled out.

ANGELA LAVANDEROS: Okay, great. Thank you. At this point in time I don’t have any other questions from the audience. Do the speakers or presenters have any questions for one another?

DR. DAVID BAKER: This is Dave Baker. Just a comment for Mark. I think you raised one point toward the end that is really important. What we found is, again, our biggest quality of care problems are the patients who have dropped out of care. We have been trying to identify those patients with diabetes who haven’t gotten care as well as those that have persistently poor control and have outreach by a care manager to try and get them back into care and provide counseling and figure out what barriers they have to adequate control. Because just the reminders for the clinicians and the performance reports, they are clearly - if you are improving intermediate outcomes they need a team-based approach.

DR. MARK WEINER: Thanks, I agree with that. It also came up for us not only with diabetes, but when we were looking at patients with ED visits who hadn’t been seen. Someone suggested we should remove the patients who haven’t been seen by our practice who were having these ED visits and that seemed to be inappropriate thinking because clearly those patients really - we need to check to see if they have any source of continuing care.

DR. DAVID BAKER: Right. The other thing that you mentioned that I think is important that we have looked at actually with some - actually this is a small world. Andrea and I worked together in one of the Alliance sites we have done some work on outreach to patients with diabetes who have dropped out of care. It is very hard to get those patients in. They have established care elsewhere or they just don’t want to come in for care and I think it is important to find some way of capturing that or even giving credit to people that they have a system in place for trying to bring people back into care.

DR. MARK WEINER: Exactly. I am not a fan of quality by checkbox and you can imagine if you have a system where people just get phone calls, you have to expend some effort to really find out what is underlying the difficulty in getting them back. Is it a financial thing? Has their insurance status changed? Do they have transportation issues? You really have to work hard to figure out what is going on.

DR. DAVID BAKER: Right. For us some of the times the reason why somebody has dropped out of care is because they weren't happy with care they were given and you certainly don't want to just drop those patients out. So it's tricky.

DR. MARK WEINER: Indeed.

ANGELA LAVANDEROS: Well thank you. With no other questions, I think we are right on the 5:00 p.m. Eastern Standard Time hour so I would like to thank the speakers today for their wonderful presentations and with that we will conclude this national webinar. Thank you.

MALE MODERATOR: Thanks, Angela. On behalf of AHRQ I want to thank all of you for joining us today. Please take a moment and fill out the brief survey that came up on your screen in order to help AHRQ improve future webinars and we appreciate your time in doing that. You will be receiving an e-mail with instructions for submitting for your CME certificate. The instructions are also on the credit tab at the top of the screen. Again, thank you very much for joining us today. We hope you have a great afternoon. This concludes today's session. Take care.

END TRANSCRIPT