00001

CENTERS FOR MEDICARE AND MEDICAID SERVICES
Medicare Evidence Development & Coverage
Advisory Committee

June 17, 2009

Centers for Medicare and Medicaid Services
7500 Security Boulevard
Baltimore, Maryland

00002

Panelists

Chair
Clifford Goodman, Ph.D.

Voting Members
Saty Satya-Murti, M.D., FANN
David A. Axelrod, M.D., M.B.A.
John Cox, D.O., F.A.C.P.
Mercedes K.C. Dullum, M.D.
Mark D. Grant, M.D., M.P.H.
Mark A. Hlatky, M.D.
William H. Maisel, M.D., M.P.H.
Curtis A. Mock, M.D., M.B.A.
Joshua P. Prager, M.D., M.S.

CMS Liaison
Marcel Salive, M.D.

Industry Representative
Jose Alvir, Dr.P.H.

Executive Secretary
Maria A. Ellis

00003

TABLE OF CONTENTS

00005

PANEL PROCEEDINGS

(The meeting was called to order at 8:10 a.m., Wednesday, June 17, 2009.)

MS. ELLIS: Good morning and welcome, committee chairperson, vice chairperson, members and guests. I am Maria Ellis, the executive secretary for the Medicare Evidence Development and Coverage Advisory Committee. The committee is here today to discuss the evidence, hear presentations and public comment, and make recommendations concerning the use of Bayesian statistics to interpret evidence in making coverage decisions. The meeting will introduce Bayesian concepts, contrast Bayesian approaches with frequentist approaches, and provide some examples of using Bayesian techniques for meta-analysis. The following announcement addresses conflict of interest issues associated with this meeting and is made part of the record. The conflict of interest statutes prohibit special government employees from participating in matters that could affect their or their employer's financial interests. Each member will be asked to disclose any financial

00006

conflicts of interest during their introduction.

We ask in the interest of fairness that all persons making statements or presentations also disclose any current or previous financial involvement in a company that performs Bayesian analysis or develops guidance for the use of Bayesian analysis for public policy-making. This includes direct financial investments, consulting fees and significant institutional support. If you haven't already received a disclosure statement, they are available on the table outside of the auditorium.

We ask that all presenters please adhere to their time limit. We have numerous presenters to hear from today and a very tight agenda and, therefore, cannot allow extra time. There is a timer at the podium that you should follow. The light will begin flashing when there are two minutes remaining and then turn red when your time is up. Please note that there is a chair for the next speaker, and please proceed to that chair when it is your turn. We ask that all speakers addressing the

00007

panel please speak directly into the mic and state your names. For the record, voting members present for today's meeting are: Dr. Saty Satya-Murti, Dr. David Axelrod, Dr. John Cox, Dr. Mercedes Dullum, Dr. Mark Grant, Dr. Mark Hlatky, Dr. William Maisel, Dr. Curtis Mock, and Dr. Joshua Prager. A quorum is present and no one has been recused because of conflicts of interest. The entire panel, including nonvoting members, will participate in the voting. The voting scores will be available on our web site following the meeting. Two averages will be calculated, one for voting members and one for the entire panel. I ask that all panel members please speak directly into the mic, and you may have to move the mic since we have to share. If you require a taxicab, there is a signup sheet at the desk outside of the auditorium. Please submit your request during the lunch break. Please remember to discard your trash in the trash cans located outside of the auditorium. And lastly, and most importantly, all CMS guests attending today's meeting are only

00008

permitted in the following areas of the CMS building site: The main lobby, the auditorium, the lower level lobby and the cafeteria. Any persons found in any area other than those mentioned will be asked to leave the conference and will not be allowed back on CMS property again.

And now I would like to turn the meeting over to Dr. Barry Straube.

DR. STRAUBE: Thank you and good morning to everybody, the MedCAC panel members and also our guests from the public in the audience.

I just want to take a couple of minutes. First of all, this particular MedCAC is one of several that are a bit different than we historically have been approaching MedCAC issues, and I think it emanated when we changed the name of this committee from MCAC and added evidence development, and I think this is very very important to the evolution of what we're trying to do with the MedCAC.

Along that line, I wanted to recognize and embarrass, in the back of the room, Dr. Steve Phurrough, who I did mention at the

00009

last MedCAC, but Steve is with us today and this topic Steve came up with while he was our director of the Coverage and Analysis Group before going over to AHRQ recently. It was one of many topics that Steve and the staff thought up and have really advanced in terms of how we use the MedCAC here at CMS. I think it's only the beginning of a much larger role as time goes on as we do more comparative evidence review, cost effective analysis, et cetera.

So Steve, I want to thank you and publicly acknowledge your work while you were here for seven years, if I remember correctly, or at least five. So thank you, Steve, for this, and maybe we will dedicate this particular MedCAC to you.

Just quickly, Marcel may, this may be cutting into his remarks, but we had a little pre-brief meeting before we came up here, and I think, again, there was a lot of enthusiasm I sensed from the panel members and I appreciate that. I think the panel understands and I suspect by, actually there's more people in the audience than I anticipated, so this is a good sign too, and I think there's a number of very

00010

very important issues that we would like to see achieved from this particular MedCAC panel. One is a basic one, and it's how do we use various types of analysis, but today Bayesian analysis, in terms of interpreting evidence that we have before us, particularly in the area of coverage decision-making. But as we discussed earlier this morning, the FDA has used Bayesian analysis in their analysis, NIH uses it for a variety of things that they do. We have perhaps not used Bayesian analysis or integrated it as much into our decision-making process, at least in a formal sense. So I think how we not only use it, but how we could possibly align with FDA and NIH and other federal agencies at least, but also in some cases, are there indications where we shouldn't be aligned with them, are there good reasons why we should not be using this type or other types of analysis.

I think we've been trying to revise our coverage standards through guidance documents, et cetera, and I think this MedCAC helps us try to refine those guidance documents, making it predictable to people who

00011

want us to make national coverage decisions as to how we'll go about that process.
And then last, I think this type of analysis, not only for coverage decision-making, but again for the future, comparative effectiveness research, cost effective analysis if Congress charges us with using that in the future, when we get data from many complex sources and we're using it for other purposes that CMS tends to use it, including collecting claims and administrative data, collecting data from registries, collecting data from RCTs or observational studies, collecting data from EHRs and using it for coverage payment, quality improvement, public reporting, incentive programs and so forth, this will carry over into all those areas.
I wanted to end with acknowledging, again, that this is the first time we've had our new chair, Dr. Cliff Goodman, and our cochair, Dr. Saty Satya-Murti, here as a team, and I wanted to acknowledge both of these gentlemen for taking on these roles. It's very important to us and we think we have fantastic

00012

talent in the chair and the co-chair, as well as our panel of participants here today too. So thank you all very much. And Marcel, I turn it over to you.

DR. SALIVE: Thank you. Good morning. I'm Marcel Salive, division director in the Coverage and Analysis Group, and the designated government official for this panel. I wanted to thank the panel members, all of them individually for their willingness to serve and engaging this topic today, and I want to thank the audience for coming out, and echo the comments to Dr. Phurrough as he's bolting from the room.

Today's topic is Bayesian statistics which is, as you all know, a statistical theory and approach to data analysis that uses a method that allows us to learn from evidence as it accumulates and uses the mathematic format of Bayes theorem to combine prior information with current information on a quantity of interest. And so I think we have thought here at CMS about this, and Barry Straube outlined some of the rationale and I want to just echo that but say that really, a lot of this derives

00013

from I think the interest of people who are developing evidence using the Bayes method, Bayesian statistics, to come here and present their evidence to us and have us use it in decision-making.

And we have been hearing this desire for a number of years. I think it came much more to the forefront when FDA held their symposium and issued their guidance documents in 2006 on, the guidance that FDA issued on the use of Bayesian statistics in medical device clinical trials, and so we've heard a lot about that here at CMS. But I agree with Dr. Straube; I just would point out that there are a number of potential uses of this evidence, certainly coverage decision-making is our main focus in the coverage group, and for developing evidence and for comparative effectiveness, types of evidence. But there are probably also broader applications that we can consider as we learn today.

So with that, I want to thank you again, and we can start the proceedings. Dr. Rosemarie Hakim is going to present the questions.

00014

DR. HAKIM: Hi, and welcome to the MedCAC. As everyone has said, this is a little bit unusual for us, and we have a number of questions that are more theoretical than concrete and so we kind of designed them to be discussed, but we also have our traditional voting criteria.

The first voting question is: In assessing the strength of evidence for effectiveness of a medical intervention that incorporates Bayesian design or analysis, compared to a frequentist approach, discuss the following. The first thing to discuss is the greatest potential strengths in a Bayesian analysis approach, and the second is, what is the greatest potential weaknesses of a Bayesian approach.

The next question is one that we'll vote on. It's just asking for your level of confidence, asking you how confident you are that potential strengths of Bayesian approaches outweigh the potential liabilities in the design and interpretation of published studies.

The next one is another one that we're going to vote on and it's asking you, how

00015

confident are you that CMS should incorporate evidence that uses Bayesian approaches in trials or technology assessments submitted for coverage decisions, and it's asking you to think about clinical trials and technology assessment.

DR. SALIVE: Thank you. Dr. Goodman.

DR. C. GOODMAN: Yes. Just a question for Maria Ellis. Would you like us to introduce ourselves now or later?

MS. ELLIS: If the panel could introduce themselves and disclose if they have any financial disclosure before we get started.

DR. C. GOODMAN: Okay. Cliff Goodman, vice president of the Lewin Group. I have no financial interests in companies performing Bayesian analyses. With regard to potential other conflicts, just to disclose, about a month ago I moderated a technical working section that discussed Bayesian statistics and adaptive methods in clinical trials. This was sponsored by a group called PACE with the TMVP, but was entirely a technical session.

DR. SATYA-MURTI: Saty Satya-Murti. I am a neurologist and have been a Medicare

00016

medical director for many years in the past, and I consult for some industry as well and for my own academic society, but I have no conflicts of interest for this topic.

DR. AXELROD: I'm David Axelrod, I'm a surgeon from Dartmouth. I have no financial interests to disclose and no conflicts.

DR. COX: John Cox, medical oncologist in Dallas, Texas, and I have no conflicts associated with Bayesian analysis.

DR. DULLUM: Mercedes Dullum, cardiac surgeon. I have no conflicts for Bayesian analysis.

DR. GRANT: Mark Grant, associate director at the Technology Evaluation Center, Blue Cross and Blue Shield Association, a geriatrician and epidemiologist, and have no conflicts of interest.

DR. HLATKY: Mark Hlatky, a cardiologist from Stanford, and I don't think I have a conflict, but I will say that I do consult with GE Healthcare from time to time and maybe they use this, but I don't know.

DR. MAISEL: Bill Maisel, a cardiologist at Beth Israel Deaconess Medical

00017

Center in Boston.
DR. MOCK: I am Curtis Mock, I'm a family physician and geriatrician, regional medical director with United Healthcare. I have no financial conflicts and my personal interest is improvement in patient care and outcomes.
DR. ALVIR: I'm Jose Alvir, a statistician from Pfizer. I have no conflicts with regard to Bayesian analysis.
DR. C. GOODMAN: With that, it's a delight to see Steve Goodman here to give our initial presentation. Steve, I see we're a little bit behind time, but I'm hopeful you will remain prompt, and we will give you a little warning close to when your time's up, but it's great to have you here to kick this off.
DR. S. GOODMAN: So, I'm going to talk about, very specifically it's an introduction but it's a little different introduction on Bayesian approaches in measuring the strength of evidence. And I have many, a number of colleagues here who have seen me talk on this topic in other forums. To those I apologize,

00018

but you incur an extra special responsibility of leading the audience in appropriate responses at appropriate times. So don't fall asleep.

So here we go. I'm going to start off, I heard there were voting questions, so we're going to start of with a voting question, and those who have heard this before, you are now allowed to raise your hands. So we have a well done study that's reported on a new electrical stimulator for pain control, something that MedCAC might consider. The author states that it's turned out somewhat surprisingly, in that they thought they only had a 25 percent chance of being proved before the experiment, to actually be effective in migraines, with a difference of 15 percent in the incidence of migraines in the treated and controlled group with a P of .05. So the probability that this association is real is, I'm giving you three choices, less than 75 percent, 75 to just under 95 percent, or 95 percent or above. I ask this to every audience I speak to about Bayes, so you get another two seconds. Normally I give six

00019

seconds, but we're behind time.
So with a P of .05, difference of 15 percent, how many think it's less than 75 percent? By the way, Maria is counting up the number of votes, and it has to add up to the number in the room minus the number who've seen this before. So less than 75 percent, three brave persons. 75 to just under 95 percent? Okay. 95 percent or above. So, this could really affect the validity outcome of this, it didn't quite add up to the number of people in the room.
(Laughter.)
So, the answer is less than 75 percent, which some of you probably suspected or were afraid of. And I ask this to medical audiences, I ask it to all sorts of audiences, and I will tell you that these audiences who very quickly learn not to answer any of my questions, but this one they're actually very confident on, they're not afraid of answering this question, and usually about 70 to 80 percent will answer very confidently that it's 95 percent or above.
So what's going on here? Well, I gave

00020

you a piece of information that normally we don't know how to incorporate into a question like this, which is that it had a less than 25 percent chance of being true. There is no way to incorporate that information. In fact, the very question I asked you, what's the probability that this association is real, is also not answerable using traditional techniques. So I asked you an unanswerable question using standard methods and I gave you information that was not processable using standard methods, so it's not surprising that most people would get it wrong and that they would answer that the only thing that would seem to make sense that might lead to a number like .05. So, that's why we have more to talk about.

So what were the implications of that particular sample? Well, first, that P of .05 wasn't very strong evidence, it didn't leave us very certain at the end of the day. And also, they said we don't know how to formally make use of that prior information about plausibility.

Let me say a few things I won't be

00021

saying today. I won't be saying that if we turn to Bayesian methods all our problems go away. I won't say that the only right thing for the statistics world are Bayesian and that the Bayesian approach doesn't have its own issues. I will say, however, that if we turn to Bayesian methods, that difficult issues will be discussed in the right way by the right people. I will say some of the dilemmas that CMS decision-makers face are artifacts of the statistical methods that we use in assessing the evidence that we look at and not due to the math or the scientific method, although sometimes it's shrouded in such language, and that the Bayesian perspective is the best way to think about evidence, and I'm focusing on evidence.

So here is a list of things, by the way, that have been identified as cancer risks, and I want to emphasize that these were not discovered with Bayesian methods, these were discovered with standard frequentist methods. So before we throw any rocks, if we are inclined to later, we have to remember this list. And here we go, electric razors, broken

00022

arms but only in women, fluorescent lights, allergies, breeding reindeer, being a waiter, owning a pet, fur, being short, being tall, eating hot dogs, and if you escape any of those risk categories, having a refrigerator. So we're all at risk and that is what standard methods have delivered unto us.

Here is my favorite medical journal, the New York Times. This is a very typical article, I'm sure there's something today of this sort, magnets lessen foot pain of diabetics, the study finds. As you see highlighted in the corner, a finding that runs counter to many previous studies, which if you had any sense, would have buried this to a footnote somewhere. But apparently they think this is what makes it newsworthy.

Now these are direct quotes from the article: We have no idea how or why the magnets work, but it's a real breakthrough. And while the study must be regarded as preliminary, the early results are clear and the treatment ought to be put to use immediately. So this is, again, an example of non-Bayesian thinking elevated to headlines in

00023

the New York Times.
So let's start with a little technical stuff here, and I will tell you I only have, I think, one equation in the whole thing. I try to tell the story in pictures as much as I can, because equations take an hour to explain.
So let's talk about inferences. There are many physicians and people involved in the medical field here and you understand the difference in medical inference between knowing what an illness is, a particular illness, and then knowing its symptoms. So that is what's called the deductive direction, and this is what I learned in my first two years of medical school. I was a walking encyclopedia, you give me a disease, I could give you a list of symptoms. Then I walked into the wards the third year and I wasn't told that there was a woman with Chagas disease in room four, I was told that there was a woman with a cough, rash and splenomegaly in room four and I had to go in the opposite direction, which isn't appearing here in the inductive direction, I had to go from the symptoms to the illness, and everybody here knows that that's a lot harder.

00024

So I was a walking encyclopedia, and yet I knew very very little. And that's because the inductive part, going from the symptoms to the illness is actually very very difficult to capture and is, the information that you have going in the deductive direction is not enough, and the exact same issue occurs in statistical inference, okay? So if you give me a mathematical hypothesis that is the null hypothesis, the difference as seen between two treatments is zero, I can tell you exactly what the probability is that I will observe, so this is the truth up here, and this is what we actually see, this is the results of studies. We don't actually get to see this, this is hidden in the clouds, so this is what we get to see.

And so I can tell you, if there is no difference between the treatments, exactly how probable is it that I will see a five percent difference in one direction, I'll see exactly zero difference, five percent in the other, 10 percent in the other, 15 percent, et cetera, et cetera. Those are just mathematical formulas and that's just called the deductive

00025

direction of inference. So I can tell you that if I know, how often I will see that. The problem is that that's not the business that we're in. We're shown the study, and we need to somehow divine what the underlying truth is, that's called the inductive direction, and that's what's the Bayes theorem and Bayesian inference is concerned with. I will tell you there is only one formal coherent calculus for the inductive inference and that is the Bayes theorem. There is really no controversy about that, it's a mathematical fact. This is not subject to voting by panels or whatever. What its strengths and liabilities are compared to traditional methods is another issue. Traditional rules of inference are a collection of principles and conventions to avoid errors over the long run if you know what the truth is. They don't tell us how likely our claims are to be true, so you wonder how we make any progress using traditional methods. And there's lots of good reasons and we've made plenty of good progress, and they often can be, they often make a lot of sense, but we often

00026

confront situations where they don't, and that's what this conference is about. The Bayesian theorem in words is this. We have odds that a hypothesis is true before obtaining the data, which is called the prior odds. And then we have this thing called the Bayes factor, which is what captures the evidential strength of the data, which is really going to be my focus today. Don Berry, who follows me, will not focus on that as much. And then you get this final post-study odds that the hypothesis is true, and so we will be talking more about how to use that information. So this is the one equation, and I do have to write this on the board, or in the slide. So here we have the post-study odds, that is the probability of the null hypothesis given the data, divided by the probability of the alternative hypothesis given the data. So the post-study means given the data, you have the data in hand, and what does that equal. And this is just what I showed you, it is the probability of the null over the alternative before you pull the data, so it doesn't have data here, and this is what I'm going to really

00027

focus on today. I'm not going to focus nearly as much on this, because that's where a lot of the attention to Bayes goes and there's a lot to be learned from how Bayes treats evidence, not just belief.

So over here we have this thing called the Bayes factor and it's marvelously simple. It's the probability of the data under the null hypothesis, divided by the probability of that same data under the alternative hypothesis, that is, how well the data is explained by two competing explanations. It couldn't actually be more simple, that's all it is, and we're going to explore it in much more depth.

This is what you're going to see a little bit later, I'm not going to focus on it so much, but this is the way we represent the prior over multiple hypotheses, so this might be all different degrees of treatment difference and this is the prior, either the prior belief or the prior evidence. This is the curve that represents the information from the data, which is called the likelihood function which we will talk a little bit about, and this is the probability of various

00028

treatment differences based on the combination of these two. And this curve will always sit between these two, and the way this is configured with a bell-shaped curve, it will always be a bit more certain.

And I will say right off the bat, there are many things we do with traditional methods that pretty much mimic what is done naturally under the Bayesian paradigm. In many ways they've stolen, or pretty much taken the wind out of the sails, in that you can mimic this exactly with traditional techniques of meta-analysis. So we can obviously accumulate evidence if represented in these simple ways using traditional meta-analytic techniques and come up with somewhat similar answers. It doesn't address the somewhat more complex situations and issues of design that Don will talk about. But there are many things that we do in using traditional methods. We sort of have patched them up to get some of the strengths and abilities of Bayesian methods.

So, this is another graphic that shows what Bayes is all about. We have a certain starting prior knowledge, we get data and that

00029

brings us to a new state of knowledge, final or posterior knowledge. And what this shows is that you can look at Bayes theorem from two perspectives; you can focus on this and this, you can talk about it as a calculus of belief, that is, the evidence just operates on your belief going from prior to posterior, or you can look at is as a calculus of evidence, because it tells you how to summarize the strength of the evidence, and that's going to be what I'm going to be talking about mainly, because I think that's a lot of what MedCAC does, and other bodies, they try to look at the strength of the evidence.

So, how do we know the strength of the evidence now? Well, P values are obviously at the center of that. And the man responsible for them, but also for many wonderful ideas in statistics, in fact almost everything that we use today that's non-Bayesian, is R.A. Fisher, who was a statistician, a geneticist and a polymath at the beginning of this century. What's really interesting about the history of statistics is that with the structure that he developed in the '20s, along with hypothesis

00030

testing which was proposed in the '30s, we have sort of the backbones of statistical analyses that are used in every article in the medical literature and yet this technology is, what, almost 80 years old. How many other medical technologies do we use virtually unchanged that are 80 years old? It's hard to name even one. So we should be a little bit embarrassed. It's obviously a very important foundation to build on and there are a lot of important foundational ideas, but the idea that we still use and still teach as the basic technology of physical analysis and reasoning a method that was developed at that time with actually remarkably little change should be subject to concern.

And Fisher himself was concerned. This is a graduation speech that he gave at the University of Michigan in 1958, so, you know, 30 years after he first brought forth these ideas. So, I'm quite sure it's only personal contact with the natural sciences that's capable to teach straight methodic mathematically minded people. I think it's worse in this country, the USA, than most,

00031

though I may be wrong, but certainly there is grave confusion of thought. We're quite in danger of sending highly trained intelligent young men and women out into the world with tables of erroneous numbers under their arms and a dense fog in the place where their brains ought to be. In this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there's no limit to which they could impede every sort of national effort. So what he saw was his methods being twisted in ways and used in very mechanical ways that was really an anathema to him, he was really a creative scientist, as all the statisticians were who developed these methods. And the way they've been employed and the way they were thought of when they were proposed is actually quite different. So what's the meaning of a P value? Does anybody here know, probability, plausibility, possibility? Actually, unfortunately, this is what most people think, and that's part of the problem, publish. So what is the formal definition? It's the

00032

probability of getting a result as or more extreme than the observed result if the null hypothesis, usually chance being operational, were true. Now I defy anybody to make sense out of that, okay?

And I will also note since the P value is calculated assuming the null hypothesis is true, it can't represent the probability that the null hypothesis is false. It's already assuming the truth. The question is, what does it mean about the truth, and that of course is our dilemma.

Here is the picture that corresponds to that, here's the distribution or results under the null hypothesis, here is the null hypothesis, and you see something out here that might correspond to a 10, 20, 30 percent treatment difference, and we calculate the area under that curve. And if we want to make things even more obscure, confusing, incoherent, we'll calculate it on this side and call it a two-sided P value and explain to our poor students why that makes sense.

So that's what a P value is.

Now, many, many, many people have

00033

written about the problems, I continued to be asked to write about it, I have no idea why, because there are literally hundreds of articles. One just appeared last year where it talked about 12 P value misconceptions, here is the list of 12, I'm not going to read them all to you. I tried to capture the most prevalent ones and here are the most prominent ones. The P value is not the probability of the null hypothesis, it's not the probability that you will make a type one error if you reject the null hypothesis, it's not the probability that the observed data occurred by chance, it's not the probability of the observed data under the null hypothesis. It's in fact not almost anything sensible you can think of. That's not to say it can't be interpreted with great care in reasonable ways, but it's not directly any of these things, and although I could spend hours, days explaining why, you can read that particular article or the hundreds of others that have preceded it if you want to learn why.

So here is just a little baby toy example that tells us why this P value poses

00034

such problems, and can pose problems for panels like yours, which I am sometimes a member of. So here we go, this is sort of a baby toy example that statisticians are very very fond of, but those who haven't seen these examples are always sort of flummoxed or surprised by this. Two scientists perform an experiment in which one of them applies two treatments, A and B, to the same individual, and they record which one is superior. You'll see why it takes two scientists in a minute. The precise data comes out like this, A is better, A is better, A is better, A is better, A is better, five As, and then B is better, okay? So that's our data. So the question is, what's the evidence for A being better than B? So you couldn't get a more simple example, A is, you know, each person is their own control, A is better in the first five and then B is better. So it sure looks like A is comparatively better. So you might ask, what was the design? Well, actually the reason I said there were two investigators here was because each

00035

investigator actually had a slightly different idea about what the design was. Investigator one had a controlled budget and planned to stop after six patients no matter what. They didn't clue in investigator two, who planned to stop the experiment as soon as B was preferred. So the question is, why does this matter, should it matter? Nature didn't know. The treatment didn't know. The bodies to whom it was applied didn't know. So should the evidence under these two scenarios be any different? Let's calculate the P values. So here's our design that only takes six samples, which you would call a fixed sample size design, you always look to the sample size fraction when you look for the P value for the sample size justification, and you might come up with six under a variety of samples, but here is the P value. So the P value would be the probability of what we saw plus the probability of a more extreme result, right, under the null hypothesis. The null hypothesis is that A would be preferred to B one half of the time. So what's the probability of five A preferences

00036

out of six? Well, it's one half to the fifth, because the null hypothesis says that A should be preferred one half of the time and then being preferred, and there's six possible combinations of those; it could be at the beginning, second, third, fourth, fifth, okay? And then we add the probability of more extreme data, which is that all six were A. So this comes out to .11 and we'd look at it and say it's not greater than .05, it's not significant and we need another experiment.

Let's look at the first B design. This is what might be called an adaptive design, that is, it adapts to the data in hand, and you will hear a lot more about this from Don Berry. So here we have the probability of the data that we have in hand, but here we couldn't get B coming first or B coming second, we could only get B coming at the end by definition, so the probability of that is exactly the sequence that we saw, plus the probability of getting six As and a B, seven As and a B, eight As and a B, and it turns out that that P value is .03, significant, less published.

00037

Same data. The only thing that's different is what was in the heads of the experimenters but they didn't write it down. So you might say oh, they didn't write it down. But it is a rather funny thing, that depending on what was on a piece of paper, the evidence in front of you is going to be different, the exact same evidence, that is a bit of a conundrum.

So, the conundrum is summarized too, the strength of the evidence depends on data we didn't see, that is results we didn't get, which in turn depends on what the experimenter intended to do if other data had been observed, that is, the stopping point. So the evidence exists only in the experimenter's mind. So again, if we start hearing anything about Bayes having to do with things in people's minds, let's remember this example.

So what do we do in traditional statistics? Well, we try to control this with very strict design and conduct rules, and so we know exactly what we would have done if we saw different things, that's what we try to do. So we define very carefully the set of outcomes

00038

that would have occurred under the null hypotheses because we constrain the outcome, but it's completely artificial. This is completely a demand of the method, it's not really a demand of science.

We're often at a loss, however, if those rules aren't followed exactly or if they're in any way ambiguous, and this is the problem that you confront all the time, and here is an example of such a problem being confronted in a conversation at the FDA. So this was a very well-known incident that occurred for the RIS drug Carvedilol, which Mark Hlatky probably knows quite well. It was a study design that was powered for heart failure and it was powered to look at the reduction in the symptoms of heart failure because they didn't think they would have the power to look at mortality endpoints. So they chose heart failure as the primary endpoint, not because mortality wasn't but because they didn't think they would be able to get that evidence.

Well, what happened? What happened was that there wasn't a whole lot of effect on

00039

the heart failure endpoint but there was a very big endpoint on a mortality endpoint, and this was the discussion that ensued. This is the chair of the committee: What we have to wrestle with is how to interpret P values for secondary endpoints in a trial which frankly was negative for the primary. In a trial with a positive endpoint you haven't spent all the alpha on that positive endpoint and you have some alpha to spend on the secondary endpoint. In a trial with a negative finding for the primary endpoint you have no more alpha to spend for the secondary endpoints. And then his committee members complete this spiral downward.

What are the P values needed for a secondary endpoint? Certainly we're not talking .05 anymore, you're out of this .05 stuff, and I would like to have seen what you thought was significant and at what level. What P value tells you that it's there study after study?

And Dr. Konstam says, what kind of statistical correction do you have to do to that survival data given the fact that there is

00040

no specified endpoint? I have no idea how to do that from a mathematical standpoint. Now these guys are cardiologists. What you want to hear from them is cardiology, you don't want to hear about how to spend the alpha in their pockets and how to make the adjustments, and this is where conventional methods can lead you, into this really spiral of gibberish about statistics, whereas what you should be talking about is did this make sense, what were the results of other studies, what were ancillary endpoints that support the mechanism we might propose. Those are the sorts of things that we might want to be talking about. What happened here is they had a result that sort of read against their prespecified design focus, which typically includes what is the primary endpoint and what was the secondary endpoint, and this gets into other issues as well. But you can see that it's very very difficult to constrain one's thinking when you're presented with a result that seems nonsensical. Eventually, I believe it was in

00041

another panel convened and this ended up being approved, but there are many many examples like this, where sort of a religious adherence to prespecified rules gets us into territory which we don't know how to navigate in, and then you can only default, so the Bayes factor talks a little bit about it. So, I've already defined it, it's simply the probability of data under your two competing explanations that you are considering.

So let's do the Bayes factor calculation for that little example I showed you, the five As and the B, okay? So we have here a null hypothesis, which was the probability that A preferred is a half, right? But right away we have to put something as part of the calculation which we didn't have to put with the P value, which is an alternative, we have to specify the alternative hypothesis. This doesn't come into the P value, it only comes into sample size calculations and all that, but you never hear boo about that later. So we're going to take an alternative hypothesis, the probability that A preferred is five-sixths, exactly what you saw. So I'm

00042

going to do the best case scenario for the alternative. That is, we're going to say the truth, the underlying truth is exactly what we saw, the probability that A preferred is five-sixths, okay? So let's do the calculations.

The probability of this data, five-eighths of six patients given under the null hypothesis compared to five-eighths of six patients under the alternative hypothesis, under the fixed sample size design turns out be .23. What does that mean? It means that it's about one quarter as likely under the null than under the alternative, that is, it's about four times more likely under this hypothesis than it was under this hypothesis, and that is the best case scenario we can make, in a sense the strongest case we can make against the null.

But this, there's something interesting about that. It's not nearly as low as the P value. Remember, the P value under this design was .11, so it's more than double that. So let's do the same calculation for the first B design. It turns out to be exactly the same. So the evidence is the evidence, and

00043

here, remember, the P value was .03. So there are two things we learn from this. First, this aspect of the design doesn't make a difference, the evidence speaks for itself. Second, the strength of the evidence against the null hypothesis, at least, is weaker than it was when we looked at only the P value. And the third thing is that this is very specific to the specific question we asked, which is relative to this alternative hypothesis, which is by the way the strongest supported hypothesis. And you will note that the language I used is very much the language we tend to use when we talk about evidence, it supports one thing, it supports another. We don't talk about evidence only in the negative sense which is, again, where the P value gets you.

So I've done this calculation again and changed it to an alternative hypothesis that there's only a two-thirds chance that A would be preferred out of, you know, .66 or point-eight-something, and here the evidence, again, is the same in the two situations. But now it's not quite as strong against the null

00044

hypothesis, it's weaker, and why? Because the observation is that A was preferred five-sixths of the time, so it doesn't support this alternative that it was a two-thirds chance of being preferred as strongly as it did that the truth was actually what was seen. So this hypothesis is supported a bit less strongly over the null, but still what we learned here is that the evidence is the evidence.

So, the Bayes factor doesn't depend on something that exists only in the minds of the investigators, that in some sense it's more objective than the P value, although I'd have to say the story is a little more complicated than that when we get to non-toy examples. The Bayes factor depends on what hypothesis you're comparing to the null hypothesis, so you have to be careful what question you're asking.

And I'm going to use an example of something that's closer to the kind of things that you will see, and the strength of the evidence against the null is not as strong as the P value indicates. So here's a slide that compares the properties of the two measures of evidence. The P value is noncomparative in the

00045

sense that it only is calculated relative to the null, and the Bayes factor is comparative in the sense that it compares an alternative to the evidence, or gives an alternative to the null. The P value uses observed data plus hypothetical data, that is the data you would have seen if other data and something else had happened. The Bayes factor only uses the observed data, which is why in those two scenarios the Bayes factor came out the same, it was the same observed data.

The P value doesn't use an alternative hypothesis and the Bayes method uses an alternative hypothesis that's explicit and has to be predesigned. And also, just to telegraph something I will highlight later, it's predefined and made explicit in the form of the prior. The prior is in essence a prior declaration of exactly how you're going to weight the evidence across the alternative hypothesis.

With a P value we can only talk in terms of negative evidence. With the Bayes factor we can talk about evidence being negative or positive, it supports the null, it

00046

supports the alternative, and the language and context is much more comfortable. The P value is sensitive to stopping rules, the Bayes factor is insensitive to stopping rules. The P value actually has no formal justification or interpretation, the Bayes factor has a formal justification or interpretation in the context of Bayes theorem. And it's also easy to explain, that is, it's how well what we see is explained by, you know, by our competing explanations, compare that, or it's how much our belief changes. So you can look at it through either prism. I defy any of you to try to explain what a P value is.

So understanding likelihood, I'm actually not going to spend too much time on this since in view of my pledge to avoid too much statistics they would empirically get lost, but I'm going to show them to you anyway and hope that they have some intuitive sense and you already know what they mean. This is just showing you a likelihood curve which is defined by the data or, if you observe five out of 15 events, so what the likelihood curve shows you is how much the data supports all the

00047

possible underlying hypotheses. It shows you how much the evidence supports various degrees of truth.

So what this shows you is if you observe five out of 15 events, so you have an observed cure rate of one-third, that the evidence most strongly supports, what do you know -- this is the truth down here, the true theory, so the way to read the likelihood curve is the truth along this axis, and in a sense it is the degree or the strength of the evidence for that particular underlying truth on this axis. So the strength of the evidence is strongest for the cure rate being, what do you know, one third, and then it goes down from there.

And you might be pleased to know that if you cut this at some point, you're going to get something very close to a confidence interval, the 95 percent confidence interval, and that represents cutting the curve at roughly right there. So if you cut the curve at about .15, that is evidence that's 15 percent as strong as it is for the peak value, you will get the limits of something that are

00048

very very close to the confidence interval. So this, the foundational concepts underlying Bayes evidence and some of the other traditional tools we use are very very closely connected. I will say it is not always a 95 percent confidence interval, but in simple situations it is.

Now I talked about alternative hypotheses, and I do have to use different, sort of flesh that out a bit. There's different ways you can state a hypothesis. I can state it like this, the alternative hypothesis or in any hypothesis, the cure rate is 15 percent. That's what's called a simple hypothesis, I exactly specified what the cure rate is. But I could say the cure rate is greater than 15 percent. That's what we call a composite hypothesis. Even though you see this written all the time, these sort of things, it actually represents an infinite number of hypotheses, that is, the cure rate could be 16 percent, 17 percent, 18 percent, 19 percent. So this is what's called a composite hypothesis.

A treatment difference of zero, the

00049

null hypothesis, is a simple hypothesis. If treatment is beneficial, that's a composite hypothesis, right, because it doesn't specify how much benefit. If the treatment is harmful, composite. And you will see why this is important later.

So I'm going to show you some examples from two trials, and this gets very very close to the kind of evidence we normally look at, a big RCT and a small RCT, and I'm going to ask you which provides stronger evidence against the null hypothesis. So we have our big RCT that shows a five percent mortality difference with a confidence interval from zero to 10 percent with a P of .05. The small RCT shows a 20 percent mortality difference but is very imprecise because it's a small RCT, a P of .05. The question is, what's the evidence against a null hypothesis?

What you learn when you look at this in terms of Bayesian evidence is you have to add to that question, you have to say compared to what. You can't just ask what's the evidence against the null hypothesis, because we have already been told that by this strange

00050

measure of the P value the evidence is the same, and yet, these results are dramatically different, so it doesn't make sense that these represent the same evidence against the null hypothesis.

And here are the two likely preferreds that correspond to those results. This is the big study where it peaks at the five percent level and comes down, with the confidence interval being zero to 10 percent, and here is the small study with the 20 percent peak, with the confidence interval from zero to about 40 percent. So just looking at these curves, do these represent the same evidence? They're clearly not, but how do we quantify that using Bayesian evidence? So we have to be specific about the alternative hypothesis, so let's start doing that.

So there's the degree of evidence that both curves provide for the null hypothesis, it's just that we look at the zero difference and we look at the height of the curves right there. So now let's ask a specific question. Let's ask with a Bayes factor of the null hypothesis, the difference is zero, versus the

00051

difference being five percent, the true difference under the, in the small RCT, that is this one. So what we do is we look at five, here's the five percent number, we look at the height of the curve there, and this height over that height is 40 percent. So this says that the study that showed a 20 percent difference with a confidence level of zero to 40 percent supports the null hypothesis over the hypothesis that the difference is five percent, it supports the null hypothesis 40 percent as much as this hypothesis, that's 40 percent the height of that, which we sort of know. That is, if the possible estimates are spread out all over the place, this study can't distinguish very well between a null effect and a five percent effect, that's what this is telling me. It provides very little evidence for the null versus the five percent effect. Let's ask the same question of the big RCT, what's the Bayes factor for no effect versus a five percent effect? So what we do is we extend that line up to there and we divide this by this, and that's 14 percent. That is, the five percent difference is supported one

00052

over that, about seven times more than that. And that is because this study is much better at discriminating between a five percent difference and a zero percent difference. So when we ask the question the right way, when we say compared to what and we compare it to the same thing, the evidence provided for the five percent mortality versus the null is quite different in the big RCT versus the small RCT in spite of the fact that they have the same P values.

Let's flip it around. Let's say what's the evidence for a 20 percent difference? Well, for the small RCT it's this peak over this peak, which not surprisingly is the same number we had before, .14, so the big study supports a 20 percent difference over a zero percent difference about seven times more strongly.

But let's look at the other one, and this is really interesting. The small study -- I'm sorry -- the large study which had a five percent difference, it says the Bayes factor of the null hypothesis versus the difference being as big as 20 percent in the big RCT is greater

00053

than a million, so here we have the null hypothesis is not supported very well compared to the 20 percent. Here it says that the null hypothesis is supported a million times more than the hypothesis if the true cure rate is 20 percent, and how could that be?

Well actually, this tells us just what the confidence interval tells us. The confidence interval on that big study goes from about zero to 10 percent, so it pretty much completely rules out a 20 percent change, right? That's what the large study probability tells us, that, you know, the null hypothesis is barely in the mix, 20 percent is totally out of the mix, so the null hypothesis is actually supported by this study if we're comparing it to a 20 percent difference by over a million fold. So this shows you the importance of asking the question carefully and precisely and accurately.

The only place where we get seeming equivalence of the Bayes factor is if we compare these two heights over these two heights, but these represent Bayes factors for different alternative hypotheses. This one is

00054

the delta equals five percent hypothesis, this one is the delta equals 20 percent hypothesis, and then we have Bayes factors that are equal of .14 to .14, but they're evidence for different hypotheses, and that's the problem. And the P value is the surrogate for this and this is why the P value is so confusing, it doesn't include information about the magnitudes of the effects. So when I tell you the P value is .05, which in this case is the correlate of the Bayes factor in these two, it's, while it is evidence against the null in some sense, even though the number is wrong, it's evidence against the null with respect to different alternatives and that is the problem, and that's what Bayes is very very rigorous about and that's why it makes more sense. So here we have a table that combines the data, the alternative hypotheses and the Bayes factor, and you see that we have these equivalent Bayes factors only when we have different alternative hypotheses, so it's just restating what I just said. So, I know we're technically five minutes from when I was supposed to end even

00055

though I started 20 minutes late, or ten I'm getting, okay.
DR. C. GOODMAN: Eight.
DR. S. GOODMAN: Eight, okay. I will do my best, but that was the toughest part. Suffice it to say that the P value confuses us about evidence because it doesn't take into account how large the effect is, a tiny effect in a large trial will appear to be the same evidence as a large effect in a small trial.
Now of course we do ask questions like what's the Bayes factor of the null hypothesis versus the treatment is beneficial, that is that the treatment difference is greater than zero, in which case we're talking about the evidence for the whole curve and what might we want to do then. Well, this is what Bayes does. It averages the height of that whole curve according to the pis. It compares this height compared to the average height of the rest of the curve, and that average is using the prior as a weight function. So it sort of says, what's the average evidence for benefit over the rest of the, some reasonable range.

00056

So that's what the prior is doing, that's what the Bayesian evidence measure, how it's operating.
So this is the last slide on which I'm going to spend a little bit of time with, it's sort of a Rosetta stone of translation, very very complicated, but in fact it's not quite as complicated as it looks. Here's the P value, here's the smallest Bayes factor we can muster for the null hypothesis, and you get the smallest Bayes factor when you always specify the alternative that is most supported by the data. I'll be happy to leave that there.
This is a sort of more moderate Bayesian evidence measure, but I would just focus here, and here we have words that describe the strength of the evidence. And this shows the effects of this degree of evidence, the maximum effect, the maximum effect of a P of .05 translated into the maximum -- the most powerful Bayes factor it can be translated into. It shows it's effect on various prior probabilities in the null hypothesis.
We'll focus on the P of .05. This

00057

translates into a minimum Bayes factor of .15, which is enough to bring you from a prior probability of the null hypothesis of 75 percent to a probability of 31 percent, a flip of the coin null hypothesis probability of 50 percent, down to a probability of 13 percent. That is, if you concluded that the association was real on the basis of that, you would be wrong 13 percent of the time.

And finally, if you said you were 95 percent confident that the association was real, that is that the null hypothesis only had a five percent chance, you would essentially be saying using this translation that you were only 26 percent confident that the null hypothesis were true before you started. So any statement based on the P of .05 that you're 95 percent sure that your conclusions are true means that you were 75 percent sure before you even started, at least 75 percent if not more.

And you can go down here, you see that a P of .01 corresponds to a most powerful Bayes factor, and again, a true Bayes factor under realistic conditions is going to be larger than this, gets you down from a probability on the

00058

null hypothesis of 60 percent to five percent, so you find that you have to demand somewhat more evidence looking through a Bayesian lens than you do, again, if you're looking at just the null hypothesis, that you do if you're looking at Bayesian measures of evidence, and all these numbers are higher than these numbers, but these are a lot higher. So, I can't -- I will just leave that there.

So I only start to call things, start using the word strong when I get well below .01. John Ioannidis looked at the Bayes factor in about 300 epidemiologic studies and compared them to the P values and this, the fact that there's a threshold here reflects that there's a minimum Bayes factor. But what you see here, here's the P value, here's the Bayes factor, so the P value of one percent, two percent, three percent, four percent, five percent, he did this only for significant P values. And you see here just what I've shown you, that the Bayes factors as he defined them ranged from .2, .4, .6, .8, much higher numbers. And you see that although there's this cluster around the minimum, there's a whole spread of studies

00059

that have significant P values but really very noncompelling Bayes factors, and all of these correspond to basically small studies with large effects, that's what they are.

I'll just say this, or jump to this as a famous example. A scientist gathers ten observations and discovers to her horror that they're nonsignificant, and she's quite sure there's a difference, and she comes to MedCAC and says listen, I want to show that this device or treatment works, how much more evidence can I work to provide to convince you that this actually works. And the actual answer is, using conventional approaches, start a new area of research, no amount of additional evidence can lower the overall type one error below 9.75 percent, because she's already spent her five percent alpha, and you know, in her second crack at this she's going to spend another fraction of that alpha. So there's no getting below the magic alpha of five percent if she used that threshold.

The Bayesian answer would be different. If a scientist gathers ten observations and discovers that the Bayes

00060

factor is greater than .05, what's the advice that we give about how to conduct the experiment, and the answer is keep collecting data until the Bayes factor is less than five percent, until you run out of money, time, or CMS approves.

Because in fact, even though we know that if you look multiple times at P values, you're bound to get a significant result, it's not true for Bayes factors. You can look as many times as you want, your probability of "significant" Bayes factor, and seriously we'll define it as five percent, when it by definition is less than five percent. There's a limit on how often you can get misleading evidence if you define the alternative hypothesis before you start and that's the key, defining the alternative hypothesis before you start and not wavering. So in other words, if would stop a trial -- this is just restating. This is Bayesian learning, which you will hear from Don, you'll get some more Bayesian learning from him, and that's one of the favorites.

So, a few final words on priors.

00061

Priors if informative should be evidence-based, that is, if they have a lot of information in them. Informative priors can often be represented as data equivalents, that is a prior with a 95 percent confidence interval of plus or minus 10 percent cure rate is the approximate equivalent of an RCT with 400 subjects. So you have to think if you're going to use very informative priors that you're willing to say that the evidence or belief is worth that much information. So there's not magic in Bayesian formulation, the prior does in a sense represent evidence of some sort, and if we're going to use it for public policy purposes, we should look very closely at what that evidence is.

As I've already said on many occasions, Bayes theorem is mathematically similar to a meta-analysis of the evidence in the prior to the evidence from the data.

And this is one example of, a real example of a study that we designed for kids where here was the evidence from adults, there was a lot of evidence from adults on the efficacy of a certain treatment for

00062

Guillain-Barre' disease. And in our planning we said okay, we're going to say that this represents this much evidence in kids. That is, there was evidence from about three to 400 adults that actually showed a benefit of the therapy. We said we're going to center it around one, we're not going to presume benefit, but we're going to say that the 300 adults were worth 70 kids.

Now you could take issue with that, but the discussion around how similar the disease and treatment is between adults and kids, that's a real discussion that you can have, that's a discussion I want neurologists and pathologists and doctors to have, and can inform us. It's not a discussion about spending alpha and stopping rules and things like that, and that's what I mean by saying that Bayes has the right issues discussed by the right people.

So prior specification of how we'll measure evidence with the alternative hypothesis can be seen as a prior restraint on how we measure that evidence. I really want to emphasize this. We often talk about the prior,

00063

you know, as this magic subjective nonscientific component, but it's a constraint, it's a straitjacket in the same way that the design is a prior constraint, but this is a constraint we can talk about. And the design is a prior constraint on the set of possible outcomes under the null hypothesis, and it is this constraint of the priors that sets the design free, and Don is going to speak to that. So it's critical.

Both forms of inference have constraints, but the Bayesian one has one that is more explicit and makes more sense and is subject to discussion. So what CMS needs to know, the Bayes theorem has a separable data and belief component that can be viewed as a calculus of evidence and not just belief. The likelihood-based evidence measures can have very attractive frequentist, that is error control properties, I haven't shown you that, Don will, but you don't give up using these measures of evidence. In fact you can do it just as well, you just do it along more sensible measures of evidence.

Standard inferential methods represent

00064

evidence inappropriately, use unnecessary rigidity in design and interpretation, and that the use of Bayesian evidential measures can have an impact far beyond the sometimes different numbers they produce. They affect how we talk about the evidence and who participates meaningfully in that dialogue. So, this is the entre to Don. This prior evidence defined in a broad sense should be formally incorporated in the interpretation of clinical research. It is certainly relevant to the design of clinical research. So I have a quote here from Don, we're all Bayesians in the design phase, and I couldn't find it specifically in his writings but he's repeated it I think every seven minutes between 1970 and 2009, and you're going to hear it again today, so I hope I haven't stolen his thunder. I'm going to give the final word to A.W.F. Edwards, who was a disciple of R.A. Fisher, who said this. What used to be called judgment is now called prejudice. What used to be called prejudice is now called a null hypothesis. It is dangerous nonsense dressed up as a scientific method and will cause much

00065

trouble before it's widely appreciated as such. I think that presages our meeting today. Thank you.

DR. C. GOODMAN: Thank you very much, Steve. Steve, before you leave the podium, can you distill for us in really a sentence or two --

DR. S. GOODMAN: I might say no to that.

DR. C. GOODMAN: Can you distill for us in a sentence or two from a practical standpoint, were CMS to use Bayesian interpretation in making coverage determinations based on available evidence, using those methods, that would increase the credibility of their coverage determination in exactly what way, and speak to the non-statisticians among us.

DR. S. GOODMAN: That's a very big question. I think I'm going to sound like a Supreme Court justice here or something. It really depends on the specifics and the kinds of conversation that went into the coverage decision. I will say that very often good people, sensible people looking at evidence in

00066

sophisticated ways can articulate judgments that at the end of the day are the same judgments you would make with Bayesian methods. However, it sometimes looks like you bent the rules, like issues without plausibility, and I haven't even gone into the comparative effectiveness and issues of safety, have either been broken or bent or implicit, and this sometimes makes the rationale for those decisions much more explicit. And it takes the conversations that you might have around the table around the real issues that went behind the coverage decision rather than leaving issues around studies that either deviated from their planned design, or interventions that don't have a good biologic foundation for which you might want to demand more evidence, but it's very difficult to do that under the current paradigm where you have these P .05 thresholds. So I would say that the way one articulates and incorporates formally the judgments that are made about the requirement for evidence thresholds, more of that will be on the table, a little less will be mysterious.

00067

With that said, you know, I can't say in every situation that it's going, that it will make a revolutionary difference, and if it was, then I would be impugning every decision that the MedCAC has done. That said, it allows -- I don't want to focus just on the interpretation of evidence. There's a whole domain which Don is going to talk about with Bayesian design and the kinds of studies that we set up and the kinds of studies that might be used to produce evidence to the panel that could be changed by incorporating a more liberal Bayesian approach.

So I don't want to take the question, just if we were fed a certain amount of evidence, will our decisions be better. I think this has the potential for affecting the kind of evidence that you're presented with in the first place, so it's sort of a twofold answer.

DR. C. GOODMAN: So if I were to distill what you just said, it sounds as though it could lend greater transparency to deliberation, number one. Number two, it could inform the design prospectively, CMS might be

00068

better able to inform those who would generate evidence on how to design studies that would yield more useful evidence.

DR. S. GOODMAN: Yes, and it does prevent the sometimes silly mistakes due to the adherence to sort of mechanical rules which sometimes happens, but in the hands of sensible people hopefully it doesn't; it doesn't happen too often.

DR. C. GOODMAN: Thank you. Again, before you leave, I know we may come back to further questions, but we're going to have two concise questions and even more concise answers from you, if that's possible. David first, quickly.

DR. AXELROD: I wanted to come back to your Carvedilol example that you brought up earlier, and the question of sort of secondary endpoints and subgroup analysis and that sort of stuff. And I think you made a fairly convincing argument that at least from a primary effects design phase, if you design it right, Bayesian statistics add a lot to it. Would that use of Bayesian ideas and analysis really have informed that Carvedilol

00069

discussion, because again, they didn't specify up front what that alternative hypothesis would have been so, again, how does that use of Bayesian statistics help us understand some of these sort of subgroup or secondary endpoint things that come before the group here?

DR. S. GOODMAN: That's a very complicated question actually, it's a concise question, and I will actually leave that partially to Don to answer. But the Bayesian approach to subgroup analysis and multiplicity is fundamentally different than the frequentist approach, and you can either model the relationship between surrogate endpoints and definitive endpoints, which could have been done here, where they pick one as a primary, pick one as the secondary, they're clearly related, and you could look at the relationship.

You can also look at the family of subgroup analysts and say that they will model their relationships, they will inform each other, so you can either model them as explicitly related in terms of mechanism related, or you can say that they, that you

00070

will model them as an ensemble together, and if there are outliers, you sort of pull them back, you don't believe outliers in the subgroup. But Bayesian doesn't have a magic solution to these subgroup analysis problems, but they have more coherent approaches to dealing with the problem, and Don will talk a little bit more about that in design where the issues of multiplicity are built directly in a sense a priori. But here the problem was already cast in stone by the design and by the sort of artifactual separation between secondary and primary, and it didn't capture everything that they knew about these endpoints, that was the problem. But there was a lot going on in that discussion that had nothing to do with what they knew about the relationship of these endpoints. It's actually a difficult issue.

DR. C. GOODMAN: Saty, one more and then we'll move on.

DR. SATYA-MURTI: On the adult versus child IDIG example, if I understand you correctly, that would be an example where the priors were set by clinicians who made a prior

00071

belief using their experiential data to show that so many children would be helped, correct, and that was devoid of any statistical origins, they just got together and set a prior; is that right?

DR. S. GOODMAN: That actually was set together with me and a pediatric neurologist, but the judgment call there, and it wasn't necessarily the only prior that we could have come up with, came out of judgments about how similar the treatment and the disease was in adults and children, and other evidence I didn't show you empirically showed how similar those were. So it wasn't just looking at one curve and saying okay, there's this other curve. It was using multiple sources of evidence to indicate, to relate to both the disease and the treatment in adults and children, and we had a continued conversation about that.

DR. SATYA-MURTI: And that was your alternative?

DR. S. GOODMAN: Based on evidence. I mean, there was actual empirical evidence there as a starting point. The evidence that brought

00072

us from adults to kids, though, was softer, and that's the crux of that issue.

DR. C. GOODMAN: Great. Thank you very much, Steve, we appreciate your coming and it was a splendid presentation.

Before we proceed, I know that Dr. Prager came in a few minutes later, and Dr. Prager, we need for you to introduce yourself and declare whether you have any interests.

DR. PRAGER: My name is Joshua Prager. I'm a full-time pain physician at UCLA. And I guess the closest thing I have to a conflict, which is kind of a coincidence here, is that 35 years ago I spent a full year studying Bayesian statistics at Harvard, and I think it's just a coincidence that I'm here in that regard, but I haven't done anything with it in the last 35 years, so I guess I don't have a conflict.

DR. C. GOODMAN: Thank you, Dr. Prager. I'm sure the Office of the Inspector General will have a word with you on that.

Next, we're very pleased to have Donald Berry from the Department of Biostatistics at the University of Texas. Dr. Berry, your name has been invoked at least a

00073

half a dozen times in the last hour. We would be delighted if you could condense your presentation from 60 minutes to 50, if that's possible.

DR. BERRY: Yes, sir, I'm nothing if not adaptive.

So, I do have conflicts. I jointly own with my son a company that does consulting, essentially exclusively on Bayesian statistics.

And as Steve indicated, perhaps a more important conflict, for 40 years I have been talking about this question, so I have a professional conflict, especially since for the first 30 years nobody listened. They would say things like, every time I listen to you talk, Don, I become a Bayesian for ten minutes.

So my outline is, I will tell you a little bit more about the Bayesian approach, the current use of the Bayesian approach, expanding somewhat on what Steve has so eloquently said, what is Bayesian adaptive design, predictive probabilities in design, adaptive randomization including pairing drugs and biomarkers, the way we have to go in medical research and drug development for

00074

example. And I'll tell you about I-SPY2, which is a joint venture of the NCI, the FDA, a consortium called the Foundation for the NIH, which is funded by drug companies. Adaptive dose finding, clinical utility. And CISNET. CISNET is Cancer Intervention Surveillance and Network and this is something that was funded by the NIH, it was seven models addressing the question of breast cancer mortality reduction in the United States, what was the cause, was it treatment, was it screening, what combination, one of those models was Bayesian, and I will tell you about the use and the role.

Practical advantages of Bayes, online learning, Steve talked a little bit about that, I'll tell you a bit more.

Predictive probabilities. If there's anything -- I mean, Steve indicated that there's some things that frequentists can't do; counter-frequentists can do essentially anything. The problem is that you have to be, you have to go outside of your philosophy in order to do the clever things. Probably the hardest thing that frequentists have to do is

00075

predicting where they're going. They can say if you know the parameter, if I can do a prediction on the null hypothesis, I can do predictions, I can say what the probability distribution is of the future results, but the parameters take it to be known.

It's only the Bayesian who can say on the basis of what I know today, the parameter has itself a distribution. I can put those two pieces of information, the future uncertainty and the current uncertainty together to talk about what is the probability distribution of the future results given where I am today. And that helps in monitoring trials, that helps in building trials that are efficient, and I will give you some examples of that.

Modeling, of course, all statisticians can do modeling, just an empirical observation, and I guess all of them are empirical.

Bayesians do more modeling. You will hear about some today, hierarchical modeling, for example, longitudinal modeling in cancer, but generally in medicine we err in looking at endpoints that are different from one study to another. So some are early endpoints in

00076

cancer, in Phase II we look at tumor response or progression-free survival, in Phase III we look at overall survival, and never the twain shall meet.

We've got to be modeling what happens to an individual patient over time and looking at things like tumor response and progression and survival, but also biomarkers that come into the equation, biomarkers could be thinking more of a standard, like MRI, for example, but also the various 'omics part of that revolution and decision analysis, another practical advantage of Bayes.

All right. This is Bayesian adaptive science. At M.D. Anderson, my home institution since I got there ten years ago, we have run over 300 trials from this perspective, I think that's probably more than the rest of the world combined. Most of them were Phase I and Phase II trials. As Dr. Salive indicated, the Center of Devices at the FDA about 12 years ago initiated a Bayesian approach following a mandate from Congress to do things which are so-called least burdensome, and they said oh, Bayes. And recently in 2006, as he indicated,

00077

with Bayesian guidance in the past 12 years there have been over 20 PMAs that have been approved and maybe five 10-Ks.

All of our drug companies are dabbling at least in the area, doing Bayesian adaptive designs and sometimes niche drugs. Other companies such as Eli Lily, Wyeth, Novartis, it's a substantial part of their portfolio that they use in the early phases, adaptive Bayesian design, and sometimes in Phase III trials and hopefully more in Phase IV trials, which is more the interest of CMS.

Some Bayesian device applications, and I just have this for you to look through, some areas in which the Bayesian approach has been used in drugs. And just to contrast oncology and migraine, in migraine the registration endpoint is two hours pain-free, and so the information that's available is essentially instantaneous.

There's a small matter of the logistics and the information flow, but -- I'm joking about a small matter, it's not a small matter -- but these things have been conquered by many CROs, and you can get the data almost

00078

instantaneously as to what the results are. It's not clean results, it's not audited results, it's not results that contain all of the lab values, but it's the information that we use, then, to say well, okay, what dose do we want to use for the next patient that comes in, for example, or are we done yet. In oncology and many of these other diseases, Alzheimer's, lupus, obesity, the information is not immediate. And if it's overall survival, for example, in cancer, even if it's something as horrible as pancreatic cancer, it's many months and maybe years before the information accrues, and so that raises the issue and need for doing longitudinal modeling. These are in order, my understanding of the way in which adaptive designs are used in the world of drug development, medical device development. The most common is the early stopping, historically stopping for efficacy, more recently and I think very importantly, stopping for futility. Dose finding is the second most common use of Bayesian adaptive methods. Seamless phases, where, you know, this notion that after a phase

00079

of drug development we know the answer to some question, we know the dose, you never know the dose. You have to recognize that and to move seamlessly looking at toxicity. Looking at efficacy throughout the entire course in a seamless one-two, and a seamless two-three, for example, is becoming not exactly de rigueur but is very common. Population finding, whom does my drug help, which subset of the patient population does my device work best in so that it has clinical utility? Essentially everybody is interested in this. It's an extremely difficult thing to do inferentially, scientifically, but people are doing it. Adaptive randomization, this is something that may be special for a clinical hospital. At M.D. Anderson we've done many of these trials where we base the next treatment on how well that treatment and its competitor treatments have been doing, not only historically as you heard Steve talk about in the prior presentation, but also in the trial. So if a treatment is doing better, we assign it with higher probability, and I will give you a

00080

couple examples of that. Ramping up accrual, and I thought this would be a biggie in the pharmaceutical world where you start out slowly, get information, get some potentially promising results, and if they're sufficiently promising, you open up to other centers. And if not, well, you don't continue in the smaller center or maybe eventually stop for futility. You know, putting in something -- I sometimes ask investigators, okay, so what is going to be the consequence of this trial? And they say -- I say suppose such and such of data and they say X. Or I say well, suppose thus and so are the data, and they say X. They don't really know why they're doing these trials. It's of course a lot better in industry, but even there putting together a coherent development within a process, not necessarily a trial but a process.

Anyway, the reason it's on the bottom of the list is I was wrong, this has not been adopted very widely. I know one trial where they're doing that in industry. The reason is that the rewards in industry are associated

00081

with accruing as fast as you possibly can, sometimes to the detriment of patients in the trial, and there are plenty of examples of that. And so it's difficult to persuade people to go, you know, at a moderate rate for a time. So just a little bit about updating because I think it's so important. Steve passed over something I wanted to show you. Consider an example, simple example, paired observations. Either the treatment does better in the pair or the control does better in the pair. So they're very simple in the sense that the null hypothesis is like tossing a coin, a fair coin, the probability of success that the treatment wins the pair is a half, and you get some data. And here are the, you know, 17 observations. The first two were successes, then failure, then a couple successes and a failure, et cetera.

The way the Bayesian approach works, as Steve indicated, is you start with a prior distribution, this is a non-informative flat prior, it's a prior that for registration purposes the FDA usually asks you to assume. It pretends that you don't know anything.

00082

After the first observation of empiric results of the success and what the mechanics are, to go from prior to posterior, you multiply the probability of the data, which depends on the parameter, which in this case is P, so you multiply by P.

Then the next observation was another success, you multiply by P again, P-squared total, and then one minus P. I'm not interested so much in the mechanics, you know, how do you go from one point to the next. What does interest me is that you can do it. You can say this is what I know today, I just made another observation and I've updated accordingly. So with every observation, you can describe what you know.

And as you know, in the frequentist approach, the evidence is based on the experiment, so you say what the experiment is going to be and you follow the experiment and then draw a conclusion, you know, do you get statistical significance or not. The Bayesian approach is much more flexible than that. So just finishing, multiply by P, another P, one minus P, et cetera. These are after ten

00083

observations; what about the 11th? The 11th observation I can tell you is going to be of two types, either the pinkish color if you get a success on the 11th pair, or the purplish color if you get a failure on the 11th pair. You know that's going to happen, either one or the other. The beauty of the Bayesian approach is you can say what the probabilities are of those happening given what we know today, and this is what we know today. So based on this, what is the probability of a failure or a success, and the answer is for those on descending, the Laplace rule of succession, there's a one-third chance of failure and a two-thirds chance of success.

Predictive probabilities are essentially monitoring trials for building good experimental designs, efficient experimental designs, and in my favorite clinical trial, as he said, we must ask where we are and whither we are tending. It applies in ordinary life and it applies in clinical trials. He didn't say, you know, it would be nice to ask where we're going, he said you must ask whither you are tending.

00084

So this is the current distribution after 17 observations and you can calculate things like the probability of the treatment being effective, the P is bigger than a half. You can calculate the predictive probability distribution in the same way that I indicated before. Or doubling the sample size, suppose if you've got 17 observations, you've got bare significance in the 17 observations. If you double the sample size, you get 34 observations, the predictive observations for the next 17 is shown here in the upper histogram.

The frequentist version of that, the frequentists will say well, assume P is equal to the maximum likely estimate, then I get this distribution. You can also assume P is equal to a half and get another distribution. All of those distributions have less variability than the right one, and it's because they incorporate the variability in the future but they don't incorporate the variability in the current understanding of what is P.

So you can calculate, for example, the probability of statistical significance after

00085

you double the sample size is 88 percent, in the case where you assume that P was equal to the maximum likely estimate, it's 96 percent, and 96 percent is woefully optimistic. Why do people do Bayesian things? Smaller trials usually, more accurate conclusions, and the objective, of course there are different objectives, but when we design a trial we say what is the theme, what are we trying to do, now let's build a trial using Bayes as a tool that does that as efficiently as possible, and one of the objectives can be treating patients in the trial as efficiently and as good as possible.

Predicting trial results, I will just let you use this, let you check this out. An important thing is that we model relationships among the various endpoints and we do simulation. Here is an example of a trial that our monitoring committee met concerning, it meets every year, or considers each trial every year. This was the neoadjuvant Herceptin. Herceptin is an antibody that targets for two positive breast cancer and the Her2 oncogene generally.

00086

And for those of you who don't know, it reverses the usual way you treat breast cancer. Usually you take out the tumor and then deliver a systemic hormone or chemo. Neoadjuvant means you leave the tumor in, deliver the systemic therapy first, and observe the tumor and the effect of the treatment on the tumor and then six months later, say, remove the tumor or where the tumor once lived and send it to pathologists, and if they can't find any tumor, that's call a pathologic complete response, and that was the endpoint of the trial.

The design of the trial was 82 patients in each group, 164 total. We met, the data monitoring committee met after 20 percent of the patients had been treated. Treatment accrual was very slow for a number of reasons. In the Herceptin arm the rate was 67 percent of 18 patients and the control arm was consistent with what we had seen previously in our institution of patients responding to this therapy. And we said, you know, in view of accrual, the rate of accrual and in view of the importance of this question -- this predated

00087

the, for those of you that know the story, this predated the four adjuvant trials totaling 12,000 patients that would be announced at some point, all being completely consistent with these results, by the way.

And we did a Bayesian probability calculation, predictive probability calculation that said that after 164 patients, if we could ever get there, which was going to be well into the future, the probability of success was 95 percent and so let's stop the trial. We did.

It was submitted to ASCO and published in the Journal of Clinical Oncology.

A purely statistical reason for the sorry performance with drugs in Phase II, and this is the usual power calculation, this is traditional powering, you know, here's the null hypothesis and this is the alternative hypothesis, this is the power that we have for detecting the alternative hypothesis. Where we get the alternative hypothesis, nobody knows. The statistician says talk to the clinician, the clinician says talk to the statistician. But this, it might be, for example, an excellent likely estimate based on frequentist

00088

data. If one were to say, what do we know about the hypothesis, it may be that that's the maximum likelihood value, but there's uncertainty associated with it. Take this uncertainty and consider the fact that maybe, you know, the truth is down here. It has some probability of being down here. And if it is down there, the power is a lot less. Maybe it's up here and if it is up here, the power is a little bit more. The concavity of this curve means that the true predictive power averaging against the uncertainty is less. My rule of thumb is for something that is 80 percent power, I automatically give them credit for 60 percent, and then they have to deserve that. So here is an example where they didn't deserve it and I will just let you look through the example. It was a stroke trial, SAINT I had been conducted, had shown an odds ratio of 1.2, so their actual likely estimate in SAINT II was 1.2, and they built an 80 percent power to detect a 1.2 as opposed to 1, and they increased the sample size in SAINT II up to 3200 in order to achieve that. In reading this paper in the New

00089

England Journal, this is, as Steve said, this is not the probability of the null hypothesis. And people were saying at the time, well, if the P value is .038, that's the probability that SAINT II is going to fail, so this is carrying it a step beyond the absurd. And if you look at, this was for modified ranking, if you look at the Barthel, if you look at the stroke impact scale and all the other things they did, they were not even close to being significant. So they advertised 80 percent, naive reduction 60 percent, but based on the other characteristics that I read in the SAINT I paper, my probability that SAINT II was going to be positive was 10 percent. The Astra Zeneca statisticians when I presented this, you know, was you're wrong, it's 80 percent. Well, I was right. I suppose I wouldn't be telling you if I weren't right, but the results from SAINT II did not meet its primary outcome, and no further development is planned. So, the morals are to do predictive power instead of power, but more importantly, build in adaptive things. I mean, this trial

00090

should have stopped for futility. When I first presented this at an NCI conference, the chairman of the data monitoring committee, Stuart Polokoff was in the audience, and he came up afterwards and said that you're exactly right, we could see that the trial was going south but we had no way that we could go about getting out of this adaptive aspect and bail out.

This is a trial just recently published last month. It was a cancer group that I design trials for breast cancer for. This was a trial looking at capecitabine versus standard therapy. The NCI said we had to do an 1800-patient trial and I said we can't, we don't have that many patients, we could do 600 and that's probably going to be enough. And they said no, you have to do an 1800-patient trial.

So I built a Bayesian predictive analysis, really a very liberal interim analysis that would stop based on a prediction that after some period of time we will know the answer, so we would stop accrual at that point. We wouldn't announce but we would stop accrual,

00091

and we advertised the trial as not being a 600-patient trial, not being an 1800-patient trial, but a trial with a sample size ranging from 600 to 1800. But lo and behold, after 600 patients we did the predictive calculation. Accrual was, you know, we frankly lied to the NCI about what the accrual would be and, you know, we were right in what we knew to be our ability to accrue these patients. And after 600 patients had accrued we did this predictive calculation, and you see that it says a Bayesian statistical design was used with the range in sample size here. Interim analyses were not of the standard type in which you cross a boundary and declare victory or not. Rather, the decision to discontinue enrollment was based on the prediction that future follow-up was likely to give a meaningful answer, and it did. And for those of you interested in seeing the results, here they are. Adaptive randomization, I mentioned that at M.D. Anderson we do a lot of trials that have this characteristic, so here's a simple three-armed trial. The PI, Francis

00092

Giles, approached me about designing a trial in AML. He said Ara-C, cytarabine is the standard therapy in this disease, I would like to take each of those arms and compare it to the experimental therapy, troxacitabine, and so it would be a three-armed trial, Phase II, and I would like to have 25 patients per arm, 75 total. And I said okay, but why don't we look at the data and if it's turning out that one of the arms is doing better than another, we'll give it higher probability, and if the probability that it's better than the others is sufficiently low, we will drop it. So he said okay.

So we built in adaptive randomization and this is the result of the trial. After the 24th patient had accrued, TI was doing sufficiently poorly that we dropped it. And after the 34th patient we stopped the trial because TA dropped. And these are the data. This is CR by date. Complete remission is important, in fact, it's a registration endpoint, and roughly speaking you don't live if you don't get complete remission. And so the standard therapy had what was quite similar

00093

to the historical rate. TI dropped after five patients with no CRs, and TA dropped after 11 patients with three CRs.

You know, we could calculate, as Steve indicated, the completely perspective design. We have calculated the false positive rate and power. Maybe we made a mistake, but if we did, it wasn't a very big one. I mean, if TI is better than IA, it's not very much better. So Giles sent this to the Journal of Blood, and the editors said you can't do anything with five patients, and I wanted him to write back and say, you tell him if he gets this disease, we have a treatment for him. But he's nicer than I am, so he sent it to General Clinical Oncology, and they said it was clear that the design was a dud, but the design is wonderful so we will publish your study.

I think I'll skip this factorial design, I'll just tell you a little bit about it. You've got it in your handout, you can read. In cancer we do Phase I trials looking at toxicity, we establish an MTD, a maximum tolerated dose, and then go into Phase II. This trial design combines the two, so we start

00094

out, we walked up the dose ladder, it was a very complicated dose ladder, you'll see two dose ladders because there were two drugs and there was a schedule of dosing concurrent versus sequential so it was very complicated, like a factorial design. Only within the factorial design it's not a complete factorial, we did it adaptively, walking up such that the toxicity would allow us to walk up, but then doing this adaptive randomization stuff in the back. And so, we have a number of trials that take this tack at Anderson and this is just to show you how we do it.

I want to tell you I-SPY2, I-SPY2 is this incredibly radical idea that we can look at characteristics of patients that may be responding to a therapy, that we in fact used many therapies, there's a control therapy -- let me move forward.

This is a neoadjuvant breast cancer again, high risk, stage two or three, and the standard therapy is taxane-based. We used that, but then on top of that add either placebo or experimental agents. And the experimental agent could be one of, somewhat

00095

arbitrarily, five possibilities depending on the accrual rate. One of the issues is we were working with the Foundation for the NIH and the drug companies, and the drug companies wanted to get an answer reasonably quickly, so if the accrual is slow we're not going to be able to do five experimental arms. Drugs come along, they get inserted into the mix, so it's like a screening trial. It's like a process rather than a trial. And how big is it, it could go on forever, plugging in additional drugs. As drugs show they're either good or bad, they graduate or flunk out, and if they graduate they graduate with a diploma that says where they're good, you know, what patients are benefitting from this therapy. The primary endpoint is path CR, although we of course relate to longer-term endpoints such as survival. The surgery -- the ultimate outcome for the primary endpoint is six months, and that's reasonably rapid in cancer but it's not fast enough for us. We build in MRIs over time and look at the tumor volume, and relate the tumor volume to the ultimate endpoint, the path CR or not.

00096

And so this is something that Bayesians do just kind of naturally. They say what information do I have about the patient, what does it tell me about the ultimate outcome, and what do I have from the patients that have been treated, that are all through surgery based on what their MRI results were, and did they experience a path CR. So using all of that information, I have a current patient who has an MRI volume measured that concurs with the baseline, so what do I predict for her path CR, is it going to be a path CR or not, and what uncertainty is associated with that, and the next MRI that she gets will update that as well.

So the goal is to predict which biomarker signatures predict response to which drugs and combinations, model relationships between baseline and longitudinal markers to predict path CR. You will see that there are many biomarkers, many kinds of possibilities. False positives are rearing their heads all over the place and we have to beat them down at least to some extent. So there has to be at least some level of confirmation.

Bayesians worry about multiplicities.

00097

We graduate drugs and biomarkers to smaller more focused Phase III. Instead of having a 3000-patient Phase III trial, we have a 300-patient Phase III going. The adaptive design allows for learning, changing, adding agents over time, uses a standard biomarker. There are two kinds of biomarkers, standard and qualifying, and we're working with the FDA Center For Devices with respect to the latter. With respect to the former, the standard biomarkers have been approved and these are the ones that are used to drive treatment. We can't drive treatment off of the qualifying biomarkers. It's conceivable that the qualifying biomarkers would graduate into the standard realm where we're using it for treatment assignment, but if that happens it's an amendment to the protocol, it's not in the current protocol.

So, the FNIH was formed a long time ago, I wanted to see if I could find it and I couldn't, but the FNIH is a consortium of the NCI, this is the cancer steering committee and the FDA, and I think they said CMS, but I've been dealing with this group a lot and I

00098

haven't seen CMS there yet. But this is a consortium that includes industry as well; in fact the funding comes from industry as well as foundations.

So the control is taxane-based. We start off balancing, when a drug comes in we randomize patients to that drug based on what little information we have at the start and so it gets balanced in a randomized balance fashion. But as we get information we learn which drugs are benefitting which patients and if the probability is high enough for a particular patient, she gets that drug with a higher probability. We include combinations, possibilities for combinations.

So these are the patient strata.

There are three biomarkers. One is HER2, another is hormone receptor status, either estrogen or progesterone receptor positive, and the third is the MammaPrint, this is a 70-gene profile that has been approved by the FDA for prognosis and also prediction of response to therapy.

And so there are eight slots. This is just to give you a feeling from the previous

00099

study, I-SPY1, where the patients fell in these slides.

DR. C. GOODMAN: Don, about nine minutes.

DR. BERRY: Nine minutes, okay. These are path CR by bin, and the thing to notice from this is that there's a good deal of variability. I mean, a 17 percent probability of path CR versus 67 percent probability of CR in this portion of patients. So these are, the experimental agents are going to have to do better than these numbers or whatever the control numbers are in the context of I-SPY2. The interesting thing here for those of you who know these numbers in breast cancer is that the low numbers are the best diseases, so there is kind of a paradox here. The patients who do well, the ER positive patients, HER2 negative patients, don't benefit much from chemotherapy therapy, and that's true not only in the neoadjuvant but also in the adjuvant section.

And yet, they have a very low path CR response to therapy -- I'm sorry. And yet, they do very well -- let me start over. They

00100

do very well in the long term, they live a long time, but they don't respond very much to chemotherapy. On the other hand, HER2 positive/ER negative, the worse disease to have, except for treatments coming along to help, but the most chemosensitive, so there's a bit of a paradox.

I think I'll -- I have to get through this rapidly. It's conceivable, I mean, if I were to say -- let me back up. If I were to say dear drug company, you have a good drug for HER2 positive/HR positive, MammaPrint too, but nothing else. They say go fly a kite, there's only four percent of patients in that group. So we have to have biomarker profiles that have marketing appeal and we've reduced to like ten of them, and we calculate for those profiles, and you graduate within the profiles.

I'm going to skip to CISNET, because CISNET is something CMS may be interested in, this kind of concept. This is population modeling. This is breast cancer mortality in the United States and although we did the analysis up through 2002, it continues to drop. There was a 24 percent reduction between 1990

00101

and 2000 and the question is why, and we published a conclusion of our seven models, Effect of Screening and Adjuvant Therapy on Mortality From Breast Cancer in the United States. There were seven population models. We used common endpoints. The CDC, the NCI opened their files and let us have all of the data that they had about things like the use of mammography, who used it when, the use of things like hormonal therapy, the benefits of chemotherapy, et cetera, stage of disease over time.

And one of these models was Bayesian, guess which one. This is mammography screening over time, women ages 40 to 79, and you see that in fact it was essentially unused in the early '80s, started to come in in the mid '80s, and up until 2000 when most women had at least some screening mammograms. Adjuvant therapy over time, again increasing at about the same time. So this is a conundrum. We've got screening increasing at about the same time that therapy is increasing, are you going to be able to separate out the two.

These were simulations from our model

00102

and our model was based on fitting, so this was much more complicated than Steve was explaining. The likelihood is based on fitting the actual data. We generate a million women, give them the screening characteristics of the day depending on their age, et cetera, depending on when they had their last mammogram. When they get cancer, we give them the treatment of the day, et cetera.

And we of course don't know, from a Bayesian perspective, we don't know any of those things for sure. We don't know what the benefit is of treatment, we don't know even which treatment was given to which patient, but we incorporate that uncertainty based on parameters. We select a parameter value for all of our eight or so parameters and then generate a sample and if it agrees with the mortality that was actually observed, we accept it into our posterior distribution. And so -- and we do this again and again and again, millions of times, and of course we don't get many acceptances, we had in this case 66 simulations that were accepted, and that gave us the ability to calculate posterior

00103

distributions.
So here's one. This is tamoxifen, this is efficacy versus effectiveness. So this is from the clinical trials and it's like what Steve said about the child versus the adult. We said maybe going from a clinical trial efficacy to actual clinical use effectiveness, maybe it's not the same, so we'll discount. And the way we discounted it was quite similar to the way he did. We took the posterior distribution from the Oxford overview and inflated it by a factor of three. So this distribution is much more spread, has much more spread than does the actual data. The mean reduction in hazard of mortality or death was 28 percent, and this represents the distribution of those 66 observations just looking at tamoxifen.
And the interesting thing here, I mean, I expected that the effectiveness would not be as great as efficacy. The interesting thing is that the distribution actually shifted to the right, which, if anything, suggests that tamoxifen in actual use is more effective than in the clinical trials. Of course you will

00104

notice that the distribution is still very spread out and that reflects the fact that we don't have a great deal of information to draw this conclusion.
For example, we don't have individual women followed over time, it's all pieced together. So I'm sure a lot of times -- these are factorial runs to address what would happen if we had all, you know, everybody would get mammograms. This is apportioning the effects of the interventions. Very interestingly, we found no interaction, none of the models found an interaction between screening and therapy.
This is our 66 models. Forget those letters for a minute. And what we found was, you know, for some of the models, some of those 66, there was very little benefit from screening. This represents the uncertainty associated with the effect of screening in this direction, treatment in this direction, and these other letters are the point estimates for the other six models consistent with our model. Our model was the only one that did this variability. All the other models, because they didn't have the Bayesian approach backing

00105

them up, couldn't assess uncertainty the way that we did, but their models fit perfectly, and I suppose most anything would fit perfectly with our conclusion, but fit perfectly with the results.

So, those are the conclusions, and this was my favorite quote from CNN, statistical blitz helps pin down mammography benefits, and then the New York Times editorial, and I will stop there. Thank you.

DR. C. GOODMAN: Thank you very much, Don. Before Don is allowed to depart the podium, are there, is there a question or two that is really important right now? We will have another shot at Don later on today. Anything at this point?

Don, I will just ask you one question. You referred earlier to FDAMA and the least burdensome approach invited in that legislation, and I wonder, since 1997, I think it was, has it been borne out that indeed Bayesian approaches have contributed to least burdensome or maybe even lesser burdensome approaches, has that held up, and that might be good for us to know with regard to how it might

00106

help CMS.
DR. BERRY: So, it has. I wouldn't say least burdensome, but certainly lesser burdensome. So, we build designs for many of these companies. There was for example, and it relates to a catheter, a Biosense Webster panel meeting, a cardiology panel meeting in November where they had approached us. We built a design for them based on their slow accrual that would use this prediction, and the original study was hardwired at 250. They went to the FDA and got approval for their catheter to prevent a-fib, and with 150 patients and many of them having reached the nine-month point based on prediction and based on the early results.
In 2007, the number two medical breakthrough according to Time magazine, not that that's your -- you know -- was a sentinel node biopsy, genetic assessment of lymph nodes that we had built that, the Bayesian design stopped as soon as it was allowed to stop, and all of the hype is about the genetics, but the hype wouldn't have been in 2007, it would have been at a later time without the Bayesian

00107

design. So I think it definitely has shown lesser burdensome.

DR. C. GOODMAN: Good, thank you. With that, thank you very much, Dr. Berry, very helpful.

We are scheduled to take a 15-minute break, we're a little bit behind, and I would ask that we take, let's call it a ten-minute break, which is about as much time as it takes to get down the hall and come back. And Dr. Lewis is up next, speaking in ten minutes, and that would put us close to putting us back on time. Thank you very much.

(Recess.)

DR. C. GOODMAN: Let's reconvene, thank you for being prompt, and Dr. Lewis, the podium is yours, sir.

DR. LEWIS: Great, thank you very much. It's a pleasure to be here today. I'm speaking on behalf of the Department of Emergency Medicine at Harbor UCLA Medical Center, the David Geffen School of Medicine at UCLA, and the Los Angeles Biomedical Research Institute. In addition to my formal employers, I have a number of financial disclosures. I

00108

work as a paid consultant to Berry Consultants and as Don already mentioned, the focus of Berry Consultants is Bayesian clinical trial design and analysis, and I'm also involved as a consultant for adaptive clinical trials for a number of sponsors.

I'm going to talk a little bit about Bayesian thinking in clinical care since that was the title of the topic that was given to me. I'm going to try to clarify some questions regarding the components of the decision process since one of the key challenges facing CMS is making explicit decisions regarding coverage. I'll talk about utility functions and how they affect decisions or at least ought to affect decisions. And then I'm going to spend some time in a description of hierarchical models and how those can be used to integrate potentially heterogeneous information from multiple sources in a way that better informs the decisions that might be made. And then finally, a few closing thoughts.

In terms of examples of Bayesian thinking in clinical care, the examples are

00109

relatively sparse and most of them are quite non-quantitative. For example, in making the diagnosis of deep venous thrombosis, there are a number of clinical studies. I just grabbed one that was published back in 1997, this is not the one that's most commonly in current use, but under this system a number of risk factors for this disease and physical findings that are associated with the disease are given a point value. The points are added up and then based on the final score the patient is assigned a probability of having this disease that is qualitatively described as low probability, moderate probability or high probability.

Now in principle, this probability assessment could be used as a posterior probability if one was going to stop anyone's medical evaluation of the patient at that point, but in fact more commonly this probability assessment system is used to create a pretest probability or a prior that guides both the selection of future diagnostic tests or subsequent diagnostic tests and the interpretation of those tests.

00110

Similarly, for the diagnosis of pulmonary embolism, which is closely related to venous thrombotic disease, there are standard clinical scoring systems that are used to estimate the pretest or prior probability, and that pretest probability is used to guide the selection of tests. For example, a patient with a lower pretest probability of a serum D-dimer test may be felt to be adequate to exclude the diagnosis if the test is negative. But with a moderate or higher pretest probability, one needs to test with a higher negative predictive value or negative likelihood ratio.

For example, a CT of the chest if appropriately interpreted, in order to reduce the upper limit of the probability interval for the true probability of disease below some level that is deemed clinically acceptable, meaning there is some ill defined and often unspoken upper limit to the final post-test probability of disease that we believe is low enough, so that we feel comfortable in stating that we have clinically excluded the disease. The selection of that upper limit of the

00111

probability of disease is really based on qualitative considerations that are usually never defined and certainly aren't based on an explicit cost benefit or other decision analysis.

In terms of moving from these qualitative assessments of probability in clinical practice, which as I said are actually quite limited, there has been a desire to at least pretend that we use qualitative assessments of probability in clinical decision-making. Back in 1975, Fagan published a nomogram which essentially is a graphical method for doing a Bayesian calculation, in which the pretest odds of the disease are expressed on one axis, the likelihood ratio which is related to the Bayes factor is represented on another vertical axis, and you can use this to graphically determine the post-test probability.

So for example if one started with a pretest probability of 30 percent and the likelihood ratio for a negative test result was .2, then your post-test probability would be something around seven percent. You could have

00112

a situation in which the pretest probability was lower, say five percent, and with the same test results your post-test probability would be about one percent.

Every time I hear a lecture on evidence-based medicine, someone will bring up this slide, I stole this from someone, and I in fact have never seen this ever used in clinical practice, and I still practice about 15 hours a week clinically in the emergency department.

There are a number of reasons for this. One is the fact that defining pretest odds for an individual patient is phenomenally difficult and in fact, physicians have widely varying opinions for a single patient. But moreover, and I believe this is a key point that is poorly appreciated, and I've actually seen written, is the fact that most clinical diagnostic strategies involve the sequential application of tests whose results are likely to be correlated.

And so even though you may hear that the Bayesian approach allows sequential application of Bayes factors to update posterior probabilities, doing so requires an

00113

understanding of the correlation between those test results which virtually never exist in clinical practice.

Moving now from the question of estimating probability of diseases or probabilities of a treatment effect to the question of making decisions, how do you make a decision if you have a posterior probability distribution for the treatment effect? Well, the components of a decision problem are fourfold. The first is some sort of prior belief or prior information regarding the patient's disease state in the case of a diagnostic test or a treatment effect, and usually one also has some data or a test result to use to update that prior information to yield the posterior information, as has been well described.

But in addition, a decision problem is characterized by a set of possible actions that one might take based on that information and the goal is to make the best decision, for example in selecting and initiating the treatment for an individual patient, observing a patient without treatment, or ordering an

00114

additional diagnostic test. All three of those possible actions commonly exist in clinical practice.

The utility function, which is a key and necessary component to the decision problem, represents the value of taking a particular action when the parameter of interest, such as the presence or absence of the disease state, has a specific value. As mentioned initially in Dr. Goodman's presentation, we often don't know whether the patient has a specific disease, we know there's signs and symptoms, and hope to be able to have some probability estimates that they have a particular disease. I'm going to try to make this more concrete in a second.

The key concept in decision-making is that we select the action or the treatment that maximizes the expected utility, given our current probability or current information for the parameter of interest, for example, the true treatment effect. In this case, expected means averaged over our uncertainty in the true treatment effect or our uncertainty in the presence of a disease. And it is this use of

00115

the expected utility that characterizes the decision theoretic approach.

So as I mentioned, utility function is the value or utility of selecting a particular action, for example, a treatment or a diagnostic strategy, given a particular parameter value where one doesn't know that parameter value -- I'm sorry -- where that parameter value is assumed to be known although in fact that is rarely the case. The utility effect function should contain or should capture multiple dimensions of the benefit or harm to the patient associated with the diagnostic or therapeutic strategy given their true disease state.

There may be positive contributions, such as improvements in patient outcome both short and long term. There may be indirect benefits to the community or society through treatment of that patient, for example, through vaccination. There may be negative contributions, for example, financial costs, side effects, complications or other associated morbidity. Patient opportunity costs, the patient may require time off work in order to

00116

undergo a specific diagnostic approach or treatment. And there are provider opportunity costs, some treatments are very time consuming and labor intensive on the part of the provider, for example surgical approaches versus medical approaches.

For a utility function to make sense, all of these different contributions must be able to be expressed on a common scale. Now that is a key challenge to the use of utility functions, but it forces the different stakeholders to communicate in a common language regarding the values of their positive and negative contributions, and that's the kind of discussion that clarifies the values that are being brought to the table in making a decision and adds to the transparency of any decision that might be made, and that is a key point.

So for example, here I've illustrated a simple utility function and I just want to go through this. For example, the patient either does or does not have an epidural hematoma. An epidural hematoma is a virtual universally fatal bleeding of the arterial blood supply

00117

around the brain. For example, this is the disease that caused the death of Natasha Richardson, and it may be present or absent with the true disease state for the patient. We have a decision to make. The decision is either to obtain emergency computed tomography of the head, a diagnostic approach that was not available for Ms. Richardson, or we may not obtain that test and therefore fail to make the diagnosis and institute appropriate and rapid surgical intervention. So for example, if the disease state is that the epidural hematoma is absent, then the utility associated with obtaining a CT is a negative number because there's some cost, in this case it's largely financial cost, opportunity cost and cost associated with the radiation exposure to the patient and the incremental increase in long-term cancer associated with that test. If the epidural hematoma is absent and we do not obtain the CT scan, the utility is zero because we have incurred none of those costs. If, however, the epidural hematoma -- I'm sorry -- so if it's absent, clearly the best action to select is to not obtain the CT scan, because

00118

that maximizes the utility under the assumption that the epidural hematoma is absent. In the alternative case where the hematoma is present, getting the CT scan is associated with a utility of minus 2500, and I'm just making up these numbers, because in addition to the cost of getting the CT scan and those other costs associated with just the test, in fact the patient is going to suffer some additional morbidity and potential mortality associated with the treatment of the disease. So under this setting there is lots and lots of costs associated with the treatment.

However, if one does not obtain the CT scan and misses the diagnosis, the patient will virtually uniformly die or suffer permanent neurologic sequelae and that's associated with a very large negative utility. Under that setting the optimal action is to obtain the CT scan and to minimize the preventable morbidity and mortality to the patient. That's pretty straightforward.

But in real life we don't know until the diagnostic test is performed whether the

00119

epidural hematoma is present or absent, so we must consider the expected utilities averaged over our uncertainty in the diagnosis. So let's pretend based on the patient's mechanism of injury and additional presentation that our probability of disease is 10 percent. So in this setting before the scan is obtained, the prior probability of epidural hematoma is 10 percent, there's a 90 percent probability that the patient does not have this particular injury. In that setting to calculate the expected utility, one averages the actual utilities associated with the action and the disease over the actual probabilities that you're in either of these columns based on the presence or absence of disease.

So for example, if you have a 10 percent chance of incurring this utility if you get a CT scan, that's a negative 2500, 90 percent of that is minus 450, you add them together and your expected utility is this number. If you do not obtain the CT scan you have a 10 percent chance of incurring this utility which would give you minus 25,000, 90 percent chance of zero, and this is your

00120

expected utility. In selecting the optimal action in this setting, clearly one would select the action that has the highest expected utility or at least negative expected utility, and so you would obtain a CT scan in the case in which there is a 10 percent pretest probability, prediagnostic probability of the epidural hematoma. The point here is that the utility function clarifies exactly what it is that we're weighing in terms of the opportunity costs, the financial costs, and one can explore the ranges of pretest probability over which the best expected utility is obtained by ordering the CT scan. So again, the key concept is that we select the action that maximizes the expected utility given the current probability distribution for the parameter of interest. In the case I just gave where the uncertainty was simply a 10 percent versus a 90 percent probability of an epidural hematoma, but in the cases of considering treatment effect estimates for real clinical trials, what we usually have is a point estimate for that treatment effect.

00121

So as Don has pointed out, we should really be thinking about these point estimates in terms of the total uncertainty in the true treatment effect, uncertainty of those data, and that was the principle that led to the routine overestimation of power in a subsequent clinical trial.

This is a slide that I will not cut out because I always like it. In this case it makes certain assumptions about the expected utility associated with being assigned to the placebo group. It reflects patients' continued belief that it's always better to be in the experimental arm, but that is not borne out by the published literature.

Now I would like to move from the theoretical issue of decision-making into the consideration of heterogeneity of evidence, and this is going to touch on the challenges of integrating evidence from multiple clinical trials that may be, for example, performed in slightly different patient groups, from different patients in terms of subclasses or severity of disease, or even combining evidence regarding similar but related treatments in the

00122

same patient population. In all of these cases, patients with the same disease may be heterogeneous. There may be different comorbidities, for example, presence or absence of diabetes, hypertension, previous surgery, in terms of severity of disease or the disease subtypes. Sometimes those differences in disease subtypes are known at the time of clinical decision-making, sometimes they can only be determined later in genetic analysis. In addition, different treatments for a single disease may have characteristics in common. For example, there are classes of pharmaceutical agents that based on mechanism of action should be likely to work to similar extents in similar patients. For medical devices, for example, Fleming has a new medical device that is in terms of mechanism of action largely equivalent to current devices. In each of those cases it is naive to believe that we know nothing about the effectiveness of the treatment in one patient population or in one subclass of patients when we know quite a lot about the effectiveness in those others, and yet traditional statistical

00123

methods are extremely poor in combining that information in a way that is rigorous, verifiable and transparent. Some other clinicians borrow information; we do this informally, secretly and we never tell you about that, so we borrow information all the time. For example, if we see a patient who's different than the patients that were enrolled in a clinical trial but have some of the same characteristics, we routinely extend the apparent indications based on that clinical trial over to this new patient for a new patient population. We do that with treatment types. For example, if we have one antihypertensive that has been demonstrated to have benefit, we assume that the new hypertensive will have a similar benefit in the absence of any separate evidence. We also do this in a way that is not documented and is not quantitative. Off label use is another example of this. Some off label use is bad, some of it makes perfect sense, and the challenge is being able to tell the difference between the two. However, traditional statistical methods,

00124

frequentist statistical methods often but not always take an all or none approach to borrowing information across heterogeneous patient populations, disease categories or treatments.

In the case in which one takes an all approach in which information is just grouped from all patient populations together, think of a fixed, like a meta-analysis. This approach will fail to recognize subgroups that experience different treatment effects or complications. If one takes the no approach in which you don't allow any pooling of information from heterogeneous subgroups, you will fail to recognize situations in which there is compelling circumstantial evidence of treatment efficacy in one group, for example a group in which there is virtually no independent data and yet there's lots of data from those related groups that suggests efficacy.

We also may lead to overestimation in heterogeneity of treatment effect when we take the none approach, so the common approach in clinical trials in which we separate out the

00125

treatment effect in each of the clinically important and a priori defined subgroups may overestimate the spread of the treatment effect in those subgroups, and I'll give an example of that in a second.

So the point that I want to make about this before I get down to a picture of the specifics is that the use of hierarchical modeling and where this can be done in a frequentist way, it is much better done using the Bayesian approach and is much more transparent and understandable with the Bayesian approach. This provides a flexible method for sharing information from potentially heterogeneous groups to a degree that is justified by the consistency of information across the groups and by the limitations and the amount of information available from each group. So this allows you to share information when it's appropriate without sharing information when it is not appropriate. And this can allow us to integrate information across clinical trials, patient groups, disease categories and treatments, and this is a key technique that can be used for CMS to improve

00126

the transparency and rigor of their coverage decisions.

So let's look at the structure of our hierarchical model. In this case I'm starting the first level of my hierarchy with results from three trials labeled trial A, B and C, and in each case the trials have resulted in a single point estimate for the clinical treatment effect. Trial A appears to show some harm, trial B shows an exactly null result, and trial C, the point estimate falls to the right, demonstrating some efficacy of the treatment. So clearly looking at these trial results they're qualitatively different, and the question is how can we integrate this information to determine whether this really suggests there's a heterogeneity of treatment effect or does this demonstrate that each of the trials was too small to be convincing. In each case we need to think of the actual distribution of efficacies that are consistent with the trial results.

In the second hierarchical model we consider a hyperdistribution, which is the distribution of the treatment effects within

00127

models. Strictly speaking we are not assuming that the patients within the trials are exchangeable, we don't assume that a patient enrolled in trial A would have met the strict inclusion criteria for the patients in trial B, but we are assuming that the trial results are roughly measuring the same type of treatment effect.

This hyperdistribution, there has to be some prior information about that. We have two priors at the third level of the hierarchy, one measures the overall average treatment effect or the center of this distribution and one measures the variability, the width of this. So we have some prior information about how different we expect the average, I'm sorry, the treatment effects of the different trials to be, and it is going to be information about this hyperdistribution that tells us things both about the average effect of the treatment and about the heterogeneity of the effect across trials.

So the information from each of the three trials informs the information about the hyper distribution and what we end up with is,

00128

for each of the three trials, a new estimate of the treatment effect that takes into account not only the information from each of the trials and the strength of the evidence from each of these three trials, but also our beliefs regarding how similar the treatment effect ought to have been among the three trials.

And there's two key results here, and it's important to understand their difference. One is you get an estimate for the overall average treatment effect of this treatment in these three trials. Perhaps more importantly, you obtain three separate estimates which are better estimates of the true treatment effect within each of the trials than you would have obtained had you considered each of these three trials separately, and that is a very deep and important truth. It means that if these were different patient populations, this estimate is a better estimate for the true treatment effect in this population than the estimate that you got from that trial in isolation, and I would love to take questions about that at the end.

This is an example of the James Stein

00129

effect, and the James Stein effect works well in different trials, a lot of different trials, a lot of subgroups in trials, and it says that if the treatment works equally well in all subgroups, just naturally with statistical variation, there will be some variability in the treatment effect we observe in a clinical trial. So even if the treatment works exactly the same, there's going to be some spread in the data. And yet ironically when we look at clinical trial data and we look at the treatment effect observed in different subgroups, we take those numbers at face value without ever accounting for that excess spread which we know to occur. The James Stein estimator, which is not a Bayesian concept but can be addressed using Bayesian approaches, uses that effect to get better estimates of the treatment effect within subgroups.

So the James Stein principle says that the best estimate of the true treatment effect in a subgroup of patients within a clinical trial is not the treatment effect observed in that subgroup if there are three or more subgroups. And this is a statement that unless

00130

you thought about it for a long time, it ought to bother you. If this does not bother you, then I'm not explaining it well. So here's an example of this. This is a clinical trial performed at my institution and it is a comparison of bag valve mask ventilation or using a mask and resuscitation bag versus endotracheal tube into the trachea. In the out of hospital, so this is a paramedic setting for critically ill or injured children, this was conducted in Los Angeles and Orange County, it was sponsored by what was then AHCPR and a number of other federal agencies. It was a relatively large study, especially for treating critically ill and injured children in the prehospital study, there were 830 children. It was published in JAMA in about 2000. So we compared endotracheal intubation to bag valve mask ventilation and the primary outcome was survival to hospital discharge among these critically ill and injured children. The overall results if you just pooled all of the different subgroups of patients in terms of their indication for supplying airway intervention was that there

00131

was no improvement in survival to hospital discharge. However, a number of clinicians felt very strongly that different subgroups of patients, for example children suffering from poor respiratory arrest or from near drowning, would be more likely to benefit from an airway intervention since the cause of their severe illness was primarily respiratory in nature. So for example, this would distinguish them from patients who suffered multiple trauma in which the initial insult was not primarily respiratory in nature. And this is what we actually got from the trial. So on this axis I have the estimated odds ratio, so these are just point estimates, and I have not included the uncertainty which is quite broad for a number of subgroups of patients which were defined a priori in the protocol, so patients with multiple trauma, traumatic brain injury, near drowning, cardiopulmonary arrest, SIDS, physical abuse, respiratory arrest. And so these numbers are based on the typical calculation of the odds ratio based on patients

00132

only within each of the individual subgroups. These data are slightly different from the published data because the published data allowed the patients to be members of multiple subgroups and I couldn't do this, otherwise I'd be double counting patients. If I instead use a hierarchical Bayesian model, and for the purists here I just used an empirical Bayes approach just to make the point, I obtain new estimates for the true treatment effect in each of the subgroups that are much more tightly clustered around the average treatment effect, because this corrects for the James Stein effect and gives me new improved estimates. The irony here, and this should be disturbing to anybody who looks at subgroups in clinical trials, is that these are the numbers that we report for the estimated treatment effect, this is what's in the JAMA publication, and these are in fact the best estimates. So patients in whom it appeared there was evidence of harm or benefit, in fact if you revise the estimates to take into account this effect, will show that there's essentially no estimates of any benefit of this therapy in the

00133

prehospital setting and there's some residual benefit of harm with weak evidence in one group.

So that just shows how for a published clinical trial, how the use of a hierarchical model can qualitatively change the conclusions you would draw regarding the likely treatment effect in clinically important subgroups. But this use of hierarchical models can be used not just in the setting of subgroups, it can be used to integrate information either across clinical trials, that was the first set of graphs I showed, patient groups, that was the subgroups, or even these categories or treatments, which should be a key interest in informing coverage decisions for CMS.

So you can take my original graph in which I had three trials, and you can simply relabel this as subgroup A, subgroup B and subgroup C, that would be similar to the results I showed for the pediatric airway trial, or you can relabel it as diseases.

So for example, if you want the best estimate of the effect of a single treatment in several different disease types that are

00134

thought to share some common mechanisms of disease and therefore the mode of action for the treatment, these estimates are actually better estimates of the treatment effect in the separate disease states than the separate trials that you had access to. So these should inform the coverage decisions rather than just these in isolation.

Once more, you can picture a situation where you initially had access to these two trials, you made a coverage decision or you made an assessment of the analysis, a new trial comes along. This allows you a seamless and transparent way to integrate that heterogeneous information, again, only to the extent that it is justified by the consistency of the data to make an updated and a more informed decision.

So, my point here is that one can use Bayesian models and hierarchical modeling specifically to integrate information from a number of sources in a way that is explicit and transparent, and continuously updatable, to aid decision-makers. But even that, the decision-making, the availability of well informed and transparent posterior probability

00135

distributions doesn't make the decision for you, it is the use of those probability distributions to calculate the expected utilities that allows you to make an informed and defensible decision. So the use of explicit expected utilities where expected means averaged over our true uncertainty using all available information will allow the evaluation of implications of different utility functions. So for example, if you make a decision regarding the use of a particular therapy and a new treatment becomes available so that the patients that don't receive the first therapy now have a secondary treatment that changes their expected outcome. Think of the epidural hematoma case for a second. The don't CT approach made the assumption that without a CT the patient would have the diagnosis remain undetected, untreated, and the natural course of the disease would ensue. If there was a secondary diagnostic approach that became available, then one could have that influence the expected outcome for the patient who forgoes the CT

00136

initially. So even with no new data on the diagnostic accuracy of the first test, an updating to the utility function could yield a new and appropriate decision in light of new information about alternative diagnostic approaches and therapies.

The use of this approach to decision-making allows the straightforward and transparent incorporation of new data, so that allows one to have a model of continuous learning, which is something that is very appealing if you don't want to redo all of your analyses every time new evidence emerges regarding each treatment. And it also allows the appropriate incorporation of all available information, for example via the hierarchical modeling.

And in the last few minutes, what I would like to do is draw a parallel between the approach of adaptive clinical trials and an adaptive and continuously learning approach to clinical decision-making or the adoption of clinical practice. So in adaptive clinical trials we use the accumulating data to help us guide the actual conduct of the clinical trial.

00137

In a sense we're using the data as a compass on how the trial ought to be conducted. So almost everybody who talks about adaptive clinical trials has a slide like this. We begin with our data collection, we analyze the data. If we don't need a predetermined stopping rule we may revise our allocation sampling rule, enroll more patients and then analyze the data. And we keep growing in this process until we meet our stopping rule, when we are sufficiently sure that we have reached a conclusion or if the trial is futile to continue. And then once we meet the stopping rule, we take our next step in the development of the drug or device.

In the clinical adoption process the way I would like to see it, we would consider the outcomes supporting a diagnostic or therapeutic approach, we analyze the available data. We ask ourselves, is there sufficient evidence to establish a standard of care. If there is evidence, we can consider that a standard of care is possible to perform this measure. If there is not sufficient evidence, we gather additional evidence or wait for

00138

additional evidence to become available. That evidence is subjected to the appropriate peer review and publication processes and when the new information is available, we can analyze that. So this is a completely analogous circular process of continual learning that helps us guide our coverage decisions in a way that is verifiable externally and transparent. In terms of the things that I've talked about up to this point, for example the hierarchical model in the use of decision functions, decision utility function, excuse me, the process of analyzing available data in my opinion should very frequently make use of updatable hierarchical models, so we use all the available information but only make pooling decisions or sharing decisions from information to the degree that is justified by the consistency and quality of the data. And then the process of deciding whether the evidence is sufficient to establish a standard of care really ought to be based on a formal decision analysis in which the utilities are there out in the open for everybody to see and so that they can be

00139

updated when additional treatment or diagnostic options become available for the patients. So just a couple of closing thoughts, and I promised I would end a few minutes early. The first is that although I was given the topic of Bayesian thinking in clinical decision-making, that's in fact a very rare thing and except in a few settings, the use of Bayesian reasoning in clinical practice is qualitative and inexact at best. The utilities that drive clinical decision-making are usually ill defined, qualitative, and if one actually looks at it, physician behavior doesn't even reflect their own stated utilities. So there's a lack of coherence in clinical decision-making that is a fact of life in our current system. In terms of coverage decisions, however, poor decisions may be made when knowledge and uncertainty in that knowledge is not appropriately quantified. And the most common of this that I think we should talk about is failure to integrate information from multiple sources in a way that is flexible and justified according to the consistency of that information, and also making decisions based on

00140

utilities that have never been discussed openly, in which the multiple components of the utility function are never placed on a common scale and that are not public and therefore open to public scrutiny.

And the last point, which I didn't put on the slide, has to do with keeping track of the quality of the information that we're using to update our posterior probabilities. Since I have a minute here, I am struck by the use as a quality measure, although no longer, of blood cultures as a quality measure for treatment of patients with community-acquired pneumonia based on the use of observational data long after it was apparent to everybody that the results of the blood culture never influenced the actual care provided to those patients. The association between blood culture, the obtaining of the blood culture and outcome of those patients was an artifact of an observational study design. The data never justified drawing the conclusion, and this begs the question, can we get back to the point where the right people are discussing the right questions.

00141

And I'll stop there. Thank you very much.
DR. C. GOODMAN: Thank you very much, Dr. Lewis. Dr. Lewis, if you would just remain at the podium for a moment or two. Yes, Mark, a question for Dr. Lewis?
DR. HLATKY: I just want to clarify this term hierarchical model, because it meant something different to me coming into this meeting than I'm hearing today, and I don't know if I'm confused or if maybe other people are confused too. My sense of it was that, you know, this was a way of analyzing data when you had like patients who were nested within doctors who were nested within hospitals or other kind of care institutions, and you needed to take account of this hierarchy of where people were. And the sense that I got from your description is totally different. Is this the same term applied to two different things or is this actually the same thing or what?
DR. LEWIS: Well, first, I'm happy to be asked a question that I can answer with, we're both right. So hierarchical models are a way of dealing with clustering of multiple

00142

different types. As you correctly stated, the clustering may be of patients within physicians, physicians within health care organizations, health care organizations within funding types.
What I was focusing on here was a situation in which the clustering was clustering of patients within clinical trials and then clinical trials were clustered within another hierarchy. So the general approach hierarchical modeling is a way of dealing with data that has a hierarchical structure in terms of the correlations and as I said, there can be multiple correlations.
So the answer is no, this is not completely different, it is exactly the same. It just has to do with what the different levels of the hierarchy represent.
DR. C. GOODMAN: Thank you.
Dr. Prager, and then we'll come back.
DR. PRAGER: I think it's well documented that physicians are highly risk averse, so that if you plug this into your utility function you're either going to come out with heavy values on the outcome or you

00143

will be altering the prior probabilities in the physician's head when making decisions. And I'm wondering, which of the two do you think that is, and a probably more important question related to what we're doing here today is how would any of this, including physician's risk aversions plugged into this model, have any applicability to the question we're asking today about the use of Bayesian thinking in CMS's whole structure?

DR. LEWIS: I have several different thoughts on that and I will try to keep it brief. The first has to do with the statement that physicians are highly risk averse. The degree to which many physicians are risk averse is based on a nonquantitative understanding of the different components of their utility function. So I tried to make the point that the utility function includes not just the cost of the treatment of the illness, or for example the likely negative utility associated with missing a diagnosis and losing a subsequent malpractice case. I happened to use a dollar amount if we missed epidural at the current cap on pain and suffering in California to try to

00144

capture that, but we don't keep track of many of these other costs, the utility costs when a patient has to miss work and those sorts of things.

So I think the first point is that for diseases that are well characterized in terms of the likely outcomes of making or missing a diagnosis, just the explicit conclusion of all of these different factors in the utility function helps clarify the factors that we ought to be balancing. I think that to the extent that coverage decisions are made based on utility functions, one can explore the range of pretest probabilities, or I'm sorry, of current probabilities of disease over which the optimal action remains unchanged.

So in my simple example I pointed out that with a pretest probability of .1, 10 percent, the optimal action was to do the computed CT scan of the head. One can look at how low the pretest probability of that disease has to be before that's no longer the optimal action, and that gives us a defensible way of defining the upper limit for the posterior probability of the disease, below which we

00145

don't have to undergo that particular diagnostic approach. So that allows a transparency and a defensibility to those limits that we currently lack.

When I asked my residents in training, how low does the probability of a pulmonary embolism have to be to not work it up, I get a number of answers that aren't based on anything other than what sounds like a small number to them. And I think that the utility functions that are developed in a publicly verifiable setting will give some credence to setting thresholds that are rational and that will yield a better allocation of our scarce resources to the diagnosis and treatment of patients.

DR. C. GOODMAN: Thank you. Dr. Dullum.

DR. DULLUM: I think you kind of answered. I was going to ask, how do you quantitate the utility function, and I think you kind of just answered that, thank you.

DR. C. GOODMAN: I think Dr. Mock was next.

DR. MOCK: I had a question along the

00146

lines of applicability of your interface between clinical and statistics. Specifically, from your discussion and your experience, do you palpably see a change in the practice leading to improved outcomes and decreased complications, and unnecessary expense in medical care? What I mean by that, I wish that you had used your DVT-PE example through the presentation, I would have loved to have seen you quantify those risks. But more specifically, if you believe that decreased variability increases efficiency, and if you use your PE-DVT, do you see it applicable to guidelines such as Milliman and Interqual where we would decrease 50 physicians treating a PE-DVT differently and have one way to treat that patient to an improved outcome and decreased complications?

DR. LEWIS: I think the answer is yes, I see the role there, but physician behavior is an explicit delineation of what the drivers of that behavior is, and those are the components of the utility function in conjunction with the willingness of either regulatory agencies, funders or specialty organizations to agree

00147

upon those utility functions so they directly lead to a threshold below which the appearing treatment or diagnostic workup is not justified by the risk of disease, and it requires both of those.

So basically somebody has to say based on our understanding of the likely outcomes of this disease, both treated and untreated, pulmonary embolism is a good example, treated it has a very low morbidity and mortality, untreated it has an extraordinarily high mortality rate. So someone has to be willing to say we believe that below a post-test probability of one percent, and that's not probably a good number, we don't think that the workup is justified, and then we have to be willing to update that as new information regarding the burdens of treatment and the alternative therapies become available.

So for example, and this is a very simple example, when we moved from having to use unfractionated Heparin by continuous drip infusion in the inpatient setting to using low molecular weight Heparin in an outpatient setting, the cost associated with the treatment

00148

went markedly down, both in terms of morbidity and the actual hospitalization costs. That should have changed our threshold of the post-test probability for initiating empiric therapy.

DR. COX: So then, we either regulate or we align incentives financially to change the behavior and the outcomes?

DR. LEWIS: Well, I would hope the first step is identifying what is a rational threshold and then aligning the incentive activities along that rational threshold.

DR. C. GOODMAN: One last question from Dr. Axelrod.

DR. AXELROD: Isn't -- throughout your presentation you talk about when you combine studies to the degree in which the data is homogenous enough you can combine those studies, and sort of inherent in that was those priors that you put up there which you sort of said, this is based on our best guesstimate of it. And I think that one of the things for those of us who don't do a lot of Bayes is that concern about, you know, there are these two big black boxes. And you know, I don't think

00149

you can address all of it but perhaps you can comment quickly on, you know, is there enough that you can kind of reassure the panel that that is not as much of a black box as it seems to be based on your presentation.

DR. LEWIS: What I was trying to show the panel is that it's possible to shine a light in that black box, and the way to shine the light in that black box is to make different assumptions regarding a third level of the hierarchy and to determine whether that affects qualitatively the decisions one would make based on the estimates of the second level of hierarchy.

So for example, when colleagues that I work with design an analysis or a clinical trial that involves a hierarchical model, it is absolutely routine to try markedly different assumptions regarding the priors at the very top level to demonstrate that with reasonable sets of data that we might expect, that would not lead to differences in the qualitative decisions regarding, for example, presence or absence of the treatment effect or the ordering of the treatment effects in terms of their

00150

relative efficacy. So you essentially do a sensitivity analysis to demonstrate that that choice is not what's driving your decision.

DR. C. GOODMAN: Thank you very much, Dr. Lewis, very helpful.

Next is Dr. Sharon-Lise Normand, who's going to address the application of Bayesian concepts in public decision-making, and I think we've already broached that topic a little bit, Doctor.

DR. NORMAND: Thank you very much for giving me the opportunity to speak today. In terms of, I have no financial interests but in terms of conflicts, I did serve on the FDA circulatory system devices advisory panel which did review some Bayesian applications. I am working with the ADHA and the ACC on updating their methodology for creating guidelines and part of that is looking at Bayesian methods to create the guidelines for evidence base. And finally, I am currently working with the FDA in the post-market surveillance setting in the Centers for Devices and Radiologic Health, and we utilize Bayesian methods.

So with that said, I'm going to talk a

00151

little bit, present you today with two problems, one is going to be a safety problem, and the other one is going to be with the idea of using a Bayesian inference to determine whether medical devices should be adopted or rejected. And in particular from a statistical standpoint, I'm going to be focusing on the use of Bayesian methods when you have sparse data, which I hope will be apparent in a second, uncertainty and heterogeneity, which Professor Lewis just spoke about, and finally function of parameters, and hopefully I will make that clear for you in a moment.

So, I'm assuming some people in this room are familiar with the following meta-analysis, and I'm grateful for the first, the previous speakers talking about meta-analyses, looking at the effect of Rosiglitazone on the risk of MI and death from cardiovascular causes, and I'm going to pick this particular meta-analysis as a starting point to demonstrate some of the issues with using a frequentist approach to meta-analysis, and I've highlighted something you can't see, but basically it's stating that there is indeed

00152

a problem, there is a safety problem with using Rosiglitazone.

So with the meta-analysis, I want to emphasize two things, one is an observational study, we didn't randomize which study could be done, so the meta-analysis is actually an observational study and we need to emphasize that, and we found that some people might not be familiar with that. And I have to thank Don and Scott Berry, because I thought they were competing with me in their earlier talks when they talked a lot about meta-analysis, but thank you anyhow.

In terms of meta-analysis to assess safety, I want to highlight the difference in the use of meta-analysis to assess safety as opposed to using a meta-analysis to assess effectiveness. And so one of the problems is, unlike effectiveness, or less so than effectiveness, the definition of safety across studies varies much more, and I think that's pretty well known. There have been some studies that actually looked at that and said indeed, you know, survival is survival, but if we're looking at a safety endpoint that's not

00153

survival, different studies have defined that in different ways, so that's a particular problem.

Therapies that increase safety risk are systematically excluded from publication, and again, that's a little bit different than the Bayesians excluding studies where there is no treatment effect finding. They are saying that even if there's a safety problem and there is a quality treatment effect, those studies are excluded anyhow. And what I want to focus on today in particular are low event rates, and I'm going to talk about the event of an MI, a heart attack.

So in some clinical trials or meta-analyses, there will be zero MIs in both treatment arms because there were two treatment arms, and sometimes there's going to be zero observed in only one of the treatment arms.

And how you handle this is critically important and typically not something that people have dealt with when you're looking at combining information for effectiveness.

So here is a picture of the data for the Rosiglitazone study and what I have on the

00154

Y axis is the event rate of heart attack by treatment arm, so on the Y axis it's the control arm, or pardon me, the Y axis is the Rosiglitazone arm, on the X axis it's the control arm. Each number represents the empiric event rate, a heart attack, in each of those studies. And the first little circle, I don't know if you can see that in red, but the reds are studies in which there is at least some -- there's some trials where there's no events in either the whole study, or one arm, and then blue is where we actually have observed events.

Now you're going to notice, there's fairly a lot of red diamonds, and what I circled are the four studies in which there were no events in both arms of the studies, because each study had a control arm and a Rosiglitazone arm. Now if I look at this, if I look in this, there were only six studies in the Rosiglitazone arm that had no events, no heart attacks, and you can see the rate of heart attacks in the control arm, in the comparator control arm, so we have six of those studies. You'll notice that we have 20 studies

00155

where there is no event in the comparator arm but there are events, that is adverse events, heart attacks in the Rosiglitazone study. So first of all you think, gee, there seem to be more studies with more events in the Rosiglitazone arm, which could be a function of a lot of things.

So first of all, you're going to see that it's difficult because you have zeroes and how do you deal with those, you can't divide by zero, how are we going to handle it. Well, typically for example, I just wanted to highlight the answer or the estimate that was reported in this particular study, and that was a NEJM article with 38 studies. Now I neglected to say that there were 42 studies overall, and so there were four studies that had no events in either arm, to get to 38, so that's apparently what the editors did, they threw away the studies where there was no event in either arm, so now we're down to 38 studies. And so they actually did find if you're going to do the P value thing that there seems to be a safety signal with the use of Rosiglitazone compared to the comparator arm,

00156

with an odds ratio of about 1.4. So how do people, and when I say people I'm going to talk about lay people, and what I mean by that is non-statisticians, typically in the past dealt with this? Well, they do a number of things. Either they drop their studies with zeroes or they add a small correction, and so what they will do is they'll add a small number less than half, and so on a two-by-two table where we've got the treatment arm, MI, no MI, the control arm, MI, no MI, and we've got some zeroes in that table, how do I fix that to actually compute an odds ratio? Sometimes what they will do is add a half to each level of that table and that way I don't have to divide by zero, I'm happy.

So that's one type of thing. You can drop the studies with zero event and then just add a half to those trials where there were only events in one arm, and that would be 38 studies. And you can see if you do that, you can see you get an odds ratio of 1.28, it's no longer statistically significant, there is no longer a meaningful P value, a signal that says there is a real problem here in terms of

00157

adverse events, meaning heart attacks. Now what other people do is, let's keep the zero events on it and add a small correction. So if I keep all 42 studies you can see I get an odds ratio of 1.26 and again, by conventional P value criteria you would say that there is no evidence of a safety problem here. Now here is the correct approach to use and it's the Bayesian approach. Now that's pretty bold of me to say, correct statistical approach, but it's the approach, it happens to be a Bayesian approach, and I'll talk a little bit about it in a second. But it's one that says okay, let's actually look at the likelihood. You've heard about the likelihood function from previous speakers, but what that does is they're able to keep all the data. I'm going to admit that in some studies I don't believe the underlying risk is zero, but I'm going to admit that indeed I can have a study where actually I will observe no heart attacks, and that's the sensible thing to do. And if you do that, you average the 42 studies, and again if you look at the P value,

00158

there's no difference, but if I could ask you to focus your attention on the odds ratio, we went from a statistically significant increased risk of an adverse event with the use of Rosiglitazone relative to comparators or controlled groups, though all the other studies showed no statistically significant P value exceeding .05. But moreover, look at the point estimates and how they change.

So what went wrong? Hopefully I can give you some clues as to what went wrong. The first thing in the paper, they said that they excluded zero total heart attacks. Well, we know that in reality, theoretically again, this is not a quote, that if you've got a binomial sample distribution which was the distribution they assumed, you can't throw away zero studies, it produces a bias. So, I don't mean to pick on this and I will tell you this is done over and over again, but in this particular study they actually did this.

Now also if you think about it, Professor Lewis talked about between study variation, and so we have, let's say 42 trials, and you could think of the 42 trials on the X

00159

axis and where we'd have the risk of the events, we have some with zeroes and some with higher values, and artificially reduced the true amount of between study variation by locking up those studies with zeroes, so again, you're artificially reducing the amount of between study variation there is. Now this actually leads to a cycle of errors when you get to that stage.

In the particular study I'm reporting on, they actually did something called a Peto odds ratio, and I suspect they adopted this type of odds ratio because it can accommodate zeroes in one arm. It can't accommodate zeroes within both arms but it can accommodate zeroes within the one arm, but in fact this is known to create bias when there's substantial differences in the control sample sizes, and in fact in these 42 studies there were huge disparities in the number of participants in the trial and control arm.

So again, they used something that was known to be biased in this setting trying to circumvent something, I think, that they felt would be a difficult to get over. And in fact,

00160

of the 42 studies, 25 percent had more in the treated than the control group. And this is why you are seeing fewer zeroes in the treated arm, because there are definitely more patients and the sample size is big enough to accommodate an observation of an adverse event. And then finally, by adding a small number to the numerator or denominator, and by that I mean adding a half to the two-by-two table, that also can cause bias and in fact it can even change the direction of the odds ratio depending on certain distributions within the tables. So you can actually go from, between the treatment and control group, going from something that is bigger than one to something that is less than one.

So, lots of different problems in reporting a very common treatment used in practice that I certainly would say caused a lot of concern from the FDA, and by the way, the FDA did their own analysis and basically agreed with these findings, again, not using sort of Bayesian approaches to deal with the fact that you can actually observe some zero events, which is very common, and for some

00161

reason they didn't take account of that.
And as I was just saying, it's not just with Rosiglitazone that this happens, so I have named a number of studies here that will be affected, such as the hemoglobin-based blood substitute that was reported in JAMA, again, adding small corrections after certain zero events. The FDA has done their own analyses of antidepressant therapies and anti-epileptic drugs, and again, they also reanalyzed the Rosiglitazone.
So I just caution here that if one, especially in the post-market studies, if we're going to look at safety in the post-market setting, that the reason people do meta-analysis is because it's too small to find a safety signal in many of these clinical trials and so it makes sense, therefore, to do a meta-analysis that can combine information across clinical trials. But when we get down to very rare events, using sort of ad hoc methods to deal with sticky problems such as zeroes, I believe that there have been some very misplaced conclusions at least based on the data in terms of looking at safety risks.

00162

Now I'm going to talk about arthroplasty, hip replacement systems. So, I talked about meta-analysis, the fact that you can do a Bayesian analysis that's going to accommodate the zeroes, it's going to reflect the uncertainty, it's going to admit to the fact that you've got between study variation. You're not going to reduce it because your frequentist method doesn't know how to handle the zeroes. And so then, that's the first part.

The second part is aligned with, I think a lot with the last piece of Professor Lewis's discussion, and that is the idea of, and I think we're going to see more and more of this, combining data from multiple and diverse data sources in order to invoke safety or comparative effectiveness.

So I'm going to talk about specifically hip replacement systems. In 2003 there were a lot of them in the U.S., about 200,000, and they were about $25,000 a pop, let's say, so it's an expensive procedure, there are a lot of them, and we have every reason to believe there will be a lot more.

00163

And why do we believe that? Because people are living longer, because there's more diabetes, there's more obesity, so we have every reason to believe that the use of these types of devices will increase.
Now what type of devices are out there? So we've got metal with polyethylene, these are all ball and socket, we've got metal on metal, and then we've got the newer ones that actually require premarket applications, sanding. So the metal on plastic, it was about, let's say a thousand that were cleared by the 510(k) path. Metal on metal, again, let's also say 150 were cleared by the 510(k) path. Now we're talking about ceramic on ceramic that were first released in the U.S. in 2003, and ten premarket applications have been approved.
Now if you think about hip replacement devices, there's a lot of information that we have short term, but what is really important is long-term consequences and effectiveness of these devices, because patients are living longer, and now you've got real estate inside your body and one really wants to get some

00164

sense of how safe, how effective are they, and we really don't have much data in the U.S. to get these devices approved, as well as other types of, let's say data that are selected in the observational phase.
So let's call this the effectiveness endpoint. That usually has been measured in most clinical trials using a short, usually one-page summary that you look at. Often one looks at survivorship and that is what is the time to hip revision that one looks at in clinical trials. And then there is a whole slew of adverse events that relate to sort of the device in and of itself that may subsequently lead to patient problems in terms of what's actually happening to them. So there could be a component where there is a breakage that causes problems for the patient, and a lot of these are going to have radiographic evidence, some of them you don't.
So again, we've got three different types of outcomes that we measure that we're interested in, and I lumped these together as adverse events, but you might want to look at those separately. What type of data do we

00165

have? Well, we have experimental data, which is the preclinical data and the typical clinical trial data, and there might be other experimental data out there in terms of non, let's say sponsored data that are out there. Now when I say preclinical data, currently the way, at least I know more about the device side, but the way the device side gets approved is that all of these paths are followed. So you've got the laboratory tests of how long the battery lasts and if it lasts for ten years, it passes; is it rusting out soon, it passes. And once it passes that hurdle then you go to another hurdle and there's information, and then you go to another hurdle. And you might have animal information and animal studies.

And once the device is passed or failed, any further evidence is completely ignored, and that is completely wrong. From a Bayesian point of view, all of that information needs to be continually integrated and updated. If we didn't think any information, if we thought information from animals were useless, we wouldn't be subjecting those poor pigs to

00166

whatever, you know, so obviously the information contained in the animal studies is helpful and so we need to integrate it. So I'm going to step further and say that we need to integrate all of the data, human data and animal data. I'm not the first person to suggest this; Bill Dumanchel suggested this in terms of looking at toxin exposures on lung disease and looked at the various mice exposure studies as well as information in people. And obviously there's a limit there, but right now it's completely forgotten about. So those are the experimental data, data in a highly controlled setting which we can use.

Then we come to observational data once it's released, but outside of -- well, it has to be once it's released because you shouldn't have access to it otherwise. So you've got FDA mandatory post-approval studies, so you've got those data which in theory will capture more complete information that would be contained in other data sources such as the Harris Hip Score, and those types of elements are important.

00167

There are some registries in the U.S. that contain information, and although we're now talking about hip replacement, the same could be true with stents, with ICDs; I just happened to take hip replacement as an example. So there are a lot of registries that one could capitalize on. Then there are registries with administrative data. So of course CMS has data in terms of the Medicare billing data, we have in-hospital billing data. So there's a lot of different data sources covering different subpopulations, and the degree of precision or completeness or breadth of those data vary by their data sources, but nevertheless, they're all informative.

And finally, there are data outside these U.S. registries and in particular for the example I'm talking about, which are hip replacements, there is actually a registry in Australia that has some pretty long follow-up in terms of these particular devices.

So, lots of different data.

So, what's the practical consideration as relates to multiple outcomes? I in the first slide talked about effectiveness,

00168

survivorship and adverse events. For some reason people treat these as well, that's one bucket, another bucket, another bucket, but that's all information about evidence for the adoption of a particular medical technology, and more importantly, a single treatment may have different effects or different outcomes, so that's the reason why there may be multiple outcomes.

Even though there's one clinical outcome, Harris Hip Score, we know that primary and secondary outcomes are always included with the outcomes, but of course the point is the different outcomes are correlated to the subject and there may be different predictors of the outcome depending on what the outcome is.

And then an important point in terms of using these very large and different data sets is the possibility of missing data, because not all outcomes are measured in every study. So if we use CMS data, we know that the Harris Hip Scores aren't there, so you could think of it as a missing data problem.

There are multiple treatments, and

00169

what I mean by that of course, we've got devices, we've got classes of devices, and you see these literally, we've got metal on metal, we've got ceramic on ceramic, we've got metal on plastic. We can think of companies, one ceramic on ceramic, or two ceramic on ceramic, so again, lots of heterogeneity. And the question is do we as a group, are you going to say okay, we are going to approve ceramic on ceramic as a device or are we going to approve company one, company two, I don't know the policy.

But in any event, you can think of these types of things, and there's also the possibility of alternative treatments and that is drugs. And so clearly in any clinical trial there would not be a suitable comparison group, and that also applies to multiple treatments. The fact that we have product synthesis, and what I mean by that is observational data. Obviously these patients aren't randomized and we've got to deal with the selection issues.

And then of course we've got multiple designs, and then we'd have to deal with

00170

cross-design synthesis. We've got randomized trials, we've got observational data. The randomized trials are studies where the individuals have been randomized but the number of studies haven't been randomized, so we've got all these problems.
We've got site effects, meaning there might be some reason to think that the outcome may vary by site, and again, it may be the hospitalization, the threshold to hospitalize somebody in Australia may be different than here. Again, I'm making that up, but there may be some reason to believe that some outcomes may vary and maybe the association of the technology might vary.
And also, of course there is the time period, over what time period are we suspecting the treatment might evolve over time.
So lots of practical considerations.
And in terms of trying to put all of these together to borrow information, to learn about outcomes that may not have been measured, or to get more precise information for the subgroups where perhaps in one data set you had much more information than in others, I really can't

00171

think of any other reasoning than to specify a full probability model, and what I mean by that is a Bayesian model.

So how do we use all of the evidence to obtain more precise evidence of safety and effectiveness of particular devices in particular patients? And again, this is related to Professor Lewis making the data sources, rather than his diagram where in one diagram it said clinical trials and another diagram said subsequent. And I'm throwing everything in together and basically saying that's all the information, how do we combine it in order to learn something.

So clearly we have to posit some mechanism that generates the observed data. And we're doing that, I'm saying we, the royal we, each investigator is doing that separately by saying this is the clinical data set we're going to posit, this is the observational data set that we're going to posit, and the animal study, I'm going to posit that it's going to give me information in order to infer something. So people are positing something, but then we have to posit some mechanism to

00172

form a whole data set. And I'm going to say that while some outcomes may be missing, we're going to assume that these outcomes are connected, and I'll show you what I mean. Now I think I have one slide with a Greek formula, and it's not meant to frighten anybody or to say that it's too complicated. I want to show you that there is lots of indices here, and the reason why the indices here are very important is because I want to enumerate the number of different sources of information. So we've got an outcome, so we've got an outcome m, which may be effectiveness, survivorship. I've got treatment k, which may be ceramic on ceramic or it may be another one. I've got a study, which could be very simple, it could have been a particular study for Medicare or a particular trial. And then we have cohort, which may be dealing with a subgroup within a trial or a Medicare cohort. And then we've got study-specific outcomes and then we've got the sampling error. The point is that the study i, cohort j, treatment k, and I've written something up there in a very loose generic sense, because

00173

I'm saying that the assumption is that the outcome, suppose it's the Harris Hip Score is greater than 70, I can model that with all this together and say that somehow it might be related, add it together to see if there is a basis for changing the treatment, et cetera. So there is a way of positing the underlying model, and the point being that even though in some studies I may not have empiric data like the Harris Hip Score and the m equals one outcome, I can use the information on those other studies to infer about, you can think about the missing outcome in the particular study I'm interested in. So we permit heterogeneity, and again, this is something that we talked about, by assuming distributions for the various components of the model, and so we can see effects due to outcome and treatment, we can see effects due to patients, blah, blah, blah. So there's lots of different effects curves that we permit; we know all of these effect curves, there's going to be some differences and heterogeneity across the various studies. So let me tell you why we should

00174

define, and again, this was motivated in the setting earlier, but let me tell you why we should be thinking about this today in 2009. Well, first of all, we now have the capability and the statistical tools, both methodologically as well as the capability, the computational capability to analyze multiple outcome measures on different scales, so lots of different people are able to simultaneously model a binary outcome, Harris Hip Score greater than 70, what's my time to hip revision, we can model those all at the same time now, which is very different than analyzing one outcome at a time, and there's lots of reasons not to model one outcome at a time, and it's mostly related to missing data. But nevertheless, we can accommodate the heterogeneity across studies and data sources, and this is a key point, that different data sources, different -- should I ignore that red light?

DR. C. GOODMAN: You've got a few more minutes.

DR. NORMAND: We can accommodate the heterogeneity across studies, we can actually

00175

combine information across studies within multiple treatment options, and we can combine now different types of studies, whether they're randomized or observational. And so the idea here is for more information from either some studies, some databases, to more precisely estimate treatment effects, and again, as I said, that led to that table in there.

So let me conclude with, what are the advantages of the Bayesian approach for quantifying the evidence? And so the first thing, it provides a coherent method for synthesizing evidence. Now that sounds like a highbrow comment, but it's very important because it makes things very transparent. Right now the model designs I don't believe in, so let's modify it. I write down the model, I know what the probability means in this setting, it's pretty straightforward.

So it's this construction of natural quantities of interest, although I didn't talk about function of parameters, I can, or we would estimate the specific class of device as particularly unsafe or we can estimate the probability that the safety risk is less than

00176

two percent, so we can actually estimate those things in a coherent framework. It does not require the modeler to do assumptions, and now you may say what, but it does not require making strong statistical assumptions, because right now if you don't combine the information, either you're doing it qualitatively in which you're not combining it, or if you are going to combine it you do need a variation that would be heroic and extremely solid method.

And also, if the studies with no events provide no information, so again, that was one assumption in the meta-analysis that I showed you, and that's actually false. So in this setting you can actually utilize studies with some zero event arms.

And it eliminates the need for approximations. Now a panel may not be so interested in technicalities, but these are quite important. And so if you have sort of a complex model and you want to combine the evidence in a coherent manner, I would have thought that an estimate at the end of the evidence analysis phase of how effective something is, I'd have to provide you some

00177

uncertainty attached to that, and if you don't use a Bayesian approach you're doing some approximations which are slippery to say the least.

So the disadvantage I have listed here is it requires more statistical knowledge and expertise to implement than standard approaches. And what I mean by that, I should be very clear what I mean by that, I know that right now almost anybody can fit a regression model by, you know, using any software package. It doesn't mean it's right, it doesn't mean you actually have interpreted the P values correctly. Again, because of the complexity, you'd actually better know what you're doing, so that's somewhat of a disadvantage that, you know, I think it's somewhat of an advantage having statisticians doing this, but in any event that's one thing you really, you know -- you need expertise in Bayesian analysis, so let me finish with that.

And finally, I would like to thank some people that I've been working with. The meta-analysis working group, we came together independently simply to formulate zero event

00178

trials and to do something virtually ad hoc with the zero event arms of the various areas that we worked in. I've looked at this from the stent side and the thrombosis, and so we looked at that. The hip replacement again, I worked with the people at the FDA in the surveillance branch, and again, I have some funding from NIH to look at combined multi-group conditions, and with that, I'll stop.

DR. C. GOODMAN: Good. Thank you very much, Dr. Normand. Can you go back to slide 20? You may need some AV help for that. Questions from the panel at this point for Dr. Normand? We have a few minutes before going to break. Yes, Dr. Prager?

DR. PRAGER: I want to thank you for really a good presentation, and I think the choice of the hip is particularly pertinent here. And give what you've been talking about, I see a whole cadre of double-edged swords that come up. And one of them is that we often, the FDA often approves a therapy, let's say in this case a hip, a specific hip replacement, without long-term data, because often they're not

00179

available. And so we don't know that if they followed it for five years and at year seven the hip completely degrades, we're left having something approved that really doesn't have a good outcome.

And so the question really comes for CMS. If we're to use a model like this and something becomes approved, how do we integrate this model into looking at outcomes after we've already been approving to go forward.

DR. NORMAND: It's a very important question, and I am not an MBA person, but I will say working with the FDA, they are revamping, at least from the devices side, post-market surveys, and so they are quite aware. So the first part of the question, you know, they are really thinking about doing the full cycle now, always updating the information now, so that's one.

The second piece of your question is, we can only approve or make decisions on the data you have available, and if you have no long-term data available, again, you want to look at as much information as you can, and you could make predictions about what could happen.

00180

But you have to have some long-term data available, and that's why we study this constant, you know, let's see what is happening in the real world. So I think this is a decision that's done at one time, so these are things that need to be looked at, and again, you can't look at everything, but there has to be some prioritization that makes sense to make sure one revisits and updates that.

DR. C. GOODMAN: Thank you.

Dr. Hlatky is next.

DR. HLATKY: Very interesting, thank you. The thing I was struck by listening to you talk about hip replacements is that there are 900, or almost a thousand different models, I guess, that have been approved under this process. And so if one is looking at, I'm assuming that, not knowing much about hip replacements being a cardiologist, that some of these devices may have device-specific problems and other ones may have stuff that's within your class. You talked specifically about the class, but I'm wondering what happens if you start seeing, I mean, how do you tease apart how much of it is, you know, this specific

00181

model is no good, versus this class is no good or this manufacturer is no good?

DR. NORMAND: Again, that relates to, although I said class in what I was saying, because I don't think we have the data for the nine studies, but pretend for the ten that have just been actually approved, there are device-specific information in those, and so you could, if that law and all that stuff permitted it, within the FDA they could look at device-specific information. So again, that's part of -- you know, right now we don't have the ability because we don't have the device-specific information, we only know that a certain type, we know it's ceramic on ceramic but we don't know if a device is made by Company K.

So the answer to the question is we'll have some information that's device-specific to look at, other data we don't. And so we're trying to borrow some information about the similarity of the devices because there is some similarity issues, and that's how we define it. Something brand new and nothing related to the past, that just doesn't happen. And so it's

00182

like a demand of, thinking would be helpful such as with, I don't know if you're familiar with the STS and the target registries where you have the device-specific names, and so it could be addressed by having more information that is device-specific.

DR. C. GOODMAN: Thank you. Yes, Dr. Dullum.

DR. DULLUM: I was thinking that this might be a benefit that CMS might look at Bayesian techniques to, once you approve a device such as the ICD, then there's always ongoing interim analysis with the possibility of disapproving it, I don't know if you ever disapprove, but which would actually be beneficial long term.

DR. NORMAND: In fact, part of the Bayes factors that were talked about earlier by Professor Goodman, where that information from those enrolled in the clinical trials could be combined with the observational registry data that CMS has mandated for collection, and you could look at it to get some Bayes factors and say here are the numbers. Somebody has to bite the bullet and say at that level it's a

00183

problem, we're going to stop, but that mechanism is definitely, we have the ability to do that, and we clearly should be doing that. We shouldn't stop, you know, again, it's always updating, updating.

DR. C. GOODMAN: Dr. Normand, with regard to the challenge that CMS might face about how to account for or embrace some messy body of evidence, you suggested in slide, I think it was slide 16, you talked about an approach to use all the evidence to obtain more precise estimates of safety and effectiveness for a particular technology, there it is. And then you go to slide 20 and you talk about having a coherent framework within which to combine this super-sized evidence.

So, do the Bayesian approaches offer a way for CMS to account for and accommodate these messy bodies of evidence where it did not have before? What's the kind of added value for Bayesian for real world mixing bodies of evidence?

DR. NORMAND: I'm going to say that the advantages are, A, there is a way to combine them using bonafide theoretically

00184

proven formulas, it is not an approximation but it's a theoretical and technical way to do it. Number two, it makes it completely transparent, and I think this is very key, this is what we're assuming, whereas in looking at analysis the way it's currently done, I think those assumptions are not transparent. I also think that it provides evidence, I think that's the key thing. Okay, you're doing things and you can only look at the data that you have, good data. But when you're doing that, you need to recognize in your statistical approach the uncertainties that are inherent in that. And I don't know of, I think the Bayesian approach is a framework that permits you to represent all of that, where it's not clear you can do that in a frequentist. Well, I could probably think of one, but that would be a special case.

DR. C. GOODMAN: But you did just say you need to know that you got good data.

DR. NORMAND: Yes.

DR. C. GOODMAN: So, does the Bayesian approach allow us to interpret the available evidence? I'm thinking in our more

00185

conventional approaches we have ways to interpret various levels of evidence, we have evidence hierarchies, we've got various things that show greater or lesser bias and so forth. Does the Bayesian toolkit allow us to make a similar or a better assessment of the quality of this evidence, since you're vouching for using almost all of it?

DR. NORMAND: While I'm vouching to use all of it, you know, if you give me data that's no good, I'm not going to use it. So there's certainly a standard level of, you know, the data elements are collected and defined appropriately. No matter what the method, you can't overcome bad data, so that's the first statement.

What it would provide you with is the type of things you want to be able to interpret, that I would claim how you're currently interpreting them is wrong. So you're placing an emphasis on P values, which you've had, you in general, in terms of concluding what evidence you have available. I think the type of the Bayesian approach in what we've talked about today is prudent. I think

00186

that everybody who speaks from a Bayesian point of view is that it provides a precise summary of the type of quantity you want, and that is, what is the evidence. And so you could combine, you do that with one data set or with 20 data sets, that's what I'm talking about. So the general Bayesian thinking is here's the evidence, these are exactly what you want and need, and that's regardless of how you combine everything. If you want to be better than all the data that you have available and it's relative good data, clean data, there is no way to combine it in a format other than the Bayesian way, because you need to adhere to, there's lots of variation rules so that you summarize it right.

DR. GOODMAN: Thank you very much. We're going to break for lunch now, but Dr. Normand, if you could keep in mind a question that might arise later in the day, which might be kind of a follow-on question, which is: Okay, let's say you want to do this. What would it take to operationalize this added facility for the Agency to undertake Bayesian approaches to evaluating evidence for coverage

00187

decisions? Thank you. This has been a great morning. I wish I could get a few college credits for it. I know it's a little bit after noon, but if we could try to reconvene at one o'clock, I know we're shaving a few minutes off for lunch, and if Dr. Sanders would be ready to ascend the stage to the podium at one o'clock, that would be wonderful. So we will see you all back at about one. Thank you very much. An enlightening morning it was, thank you very much.

(Luncheon recess.)

DR. C. GOODMAN: After our enlightening morning and a fulfilling lunch, we are going to move to a two-part presentation, starting with Gillian Sanders from Duke, and she will tag team with Don Berry once again, looking at the meta-analyses of ICDs. Dr. Sanders, if you would.

DR. SANDERS: Sure. As you heard this morning, a few years back CMS expressed an interest in exploring the advantages and disadvantages of Bayesian methods in RCTs, and particularly those in the CMS policy and

00188

decision-making arena. So in collaboration with AHRQ as partner with the Duke Evidence-Based Practice Center, we performed a systematic review of the literature and then also used a case study to explore the use of Bayesian methods in the CMS decision-making. And together with CMS and AHRQ, we chose the clinical debate of ICD therapy in the prevention of sudden cardiac death, and it's specifically the design of this case study that I'm going to be talking about today. So just a little bit of background. I'm not a Bayesian statistician. My training is a Ph.D. in medical schematics and I describe myself as a medical decision analyst, and I really focus on chronic disease modeling and then the translation of these evidence-based models into clinical practice and policy. I do, however, in those policies use Bayesian methods certainly to inform those decision models. The collaborators on this project were Lurdes Inoue, who's a matrix statistician based at the University of Washington, who actually trained with Don Berry. And then my colleagues from Duke, Dave Matchar, Greg Samsa, Shalini

00189

Kulasingam. Greg is here today and available to help with questions as well.

So, as any evidence-based practices center review, it's guided by the feedback and expertise from a technical expert panel. And so here you will see the ones that were associated with this project. On the left side are eight investigators from the clinical trials that we actually included in our analysis, and then the other side are some, four other members that represented some statistical expertise in addition.

So as many of you know, sudden cardiac death is the most common cause of death in the U.S. and it accounts for up to 350,000 deaths per year. Each year sudden cardiac death claims the lives of more people than stroke, lung cancer, breast cancer and AIDS combined. And although the overall number of cardiac deaths has decreased over the past decade, the proportion of these cardiac deaths that are sudden has increased. Of note here today, over 80 percent of sudden cardiac deaths occur in patients that are 65 years and older, highlighting particular interest to the CMS.

00190

Fortunately there's ways to prevent sudden cardiac death. The implantable cardiac defibrillator, or the ICD, is a device that monitors the heart rhythms and delivers shocks if these rhythms are detected. There's been several clinical trials on ICD therapy, and it's been demonstrated that their use can significantly reduce mortality, and it's currently the most effective therapy for preventing sudden cardiac death. The magnitude, however, of the effectiveness of the ICD in clinically identified subgroups is currently unclear.

In addition, ICD therapy is quite expensive. Current CMS reimbursement is about $30,000 per device implantation. And so although evaluations of ICD cost effectiveness by our group and by others have in general demonstrated that the ICD is a valuable use of our health care dollars, there are several researchers and policy makers that certainly looked at whether there could be ways of risk stratifying the patients further to increase the benefit and value of ICD placement.

In addition, the ICD represents a

00191

clinical domain and intervention which CMS has evaluated several times over the last two decades. Currently CMS only covers evidence development concerning these devices in the primary prevention of sudden cardiac death, which reflects really the uncertainty of several clinical policy questions.

So, this table shows some of the major ICD RCTs and their timing, each column represents a trial and -- I'm sorry, each trial is a row, and then the columns are the years in which the trial was ongoing. Those in green are considered secondary prevention trials; these are trials where the patient actually has physically experienced sudden cardiac arrest and therefore were at high risk for recurrent events.

Unfortunately, most people don't actually survive that original event, and so the latter analyses and clinical trials were really looking at what's considered primary prevention of sudden cardiac death, and these are patients that are at increased risk compared to the general population but who haven't had a previous ventricular event.

00192

In yellow here, I see there when the findings in these trials were actually made available and published, so that these could then potentially be available for subsequent trial design. For the analysis which I will discuss here today, we received access to the patient level data from eight of these trials, namely all of them except for the CIDS and the DINAMIT trial.

This table shows the clinical characteristics for the eight trials considered in our case study. There's a lot of information here so I'm just going to highlight a few things. First, note that the trials considered in the case study differed in sample size, with the smallest trial being MADIT-I, having 196 patients, and the largest being SCD-HEFT with 1,676 randomized to either ICD or control.

Moreover, there are different propositions across the trials. So for example, some trials such as CABG, MADIT-I and II, and MUSTT only have ischemic patients in their populations, while the DEFINITE trial only included nonischemic patients. The median

00193

age range was from 57 up to 65 years of age, and the ejection fraction ranged from 20 percent in the DEFINITE trial up to 45 percent in CASH. And as you can see, the distribution in the heart failure classes, it varies quite widely but overall Class IV patients are very poorly represented in the available trials. In our analysis, CMS and AHRQ were also interested in whether Bayesian methods could be used in evaluating their registry data. So we then borrowed data from the ACC-NCDR ICD registry which was formed in 2005 following CMS's coverage development for ICD therapy. The registry data collection process covered over 130 different data elements, the type of initial ICD implant, device upgrades, and then also device replacements, and the data we had access to was about 120,000 implants between January of 2005 and June of 2007, and the characteristics of the patients are represented here. Now compared to the patients that were recruited to the actual ICD trials, the registry patients are older and they actually

00194

have worse prognosis. Also, note that the registry data are only for ICD patients, that is, we don't have a control arm. And also, currently the ICD registry does not have follow-up information regarding the patient's overall survival after discharge.

And so for the purpose of illustration in our analysis here, we actually used registry data from the MUSTT study which had survival data associated with it to look at the survival comparators in the clinical trials and registry data, and I'll describe the method and issues later when I get to those results.

So as I mentioned, even with existing RCT evidence there are several clinical and policy questions that have been remaining unanswered. So this shows some of the questions that we looked at in our analysis. We looked at these both from a frequentist approach and then also using Bayesian methods. The major questions are: Are the patients within the trials similar? Is there evidence that the devices used in the different trials differ in terms of their efficacies? Is there evidence that the ICD is effective in

00195

particular patient subgroups? And can Bayesian methods be used to say anything about prognosis of patients within the ICD registry?

I will describe briefly some of the methods we used in our case study. As I mentioned, we considered patient level data from eight trials, namely MADIT-I and II, MUSTT, DEFINITE, SCD-HEFT, AVID, CASH and CABG. We used overall survival as the primary outcome, and the treatments considered were ICD with controls. The studies now seem to focus on four prognostic variables; these were age, ejection fraction, the New York Heart Association class, and the presence of ischemic disease. Now there are certainly other prognostic variables that may be as closely important, for example, the cure interval or time from MI, and we had a reviewer actually explore these additional factors, but we really wanted in this situation to explore the use of the methods.

So we performed four sets of analyses, so we used the data from individual trials, combining data from all trials, the use of registry data, and then to validate the impact

00196

we had access to aggregate versus patient level data. Given our time constraints today, I'm just going to skip over our analysis of the individual trials and instead focus on the remaining three sets of analyses.

In our analysis of data combining all trials we used both frequentist and Bayesian techniques to find data, we made adjustments for potential trial effects, adjusted for trial effects using fixed or random effects, and assuming trial-specific baseline hazard functions. Throughout our combined trials analyses we used the frequentist data as the priors we used in the Bayesian analyses. In the analysis of the registry data, we used the Bayesian methods to simulate the survival experience of hypothetical patients in a hypothetical new trial utilizing ICD and control groups in patient subgroups, and then compared the predicted and empirical survival data. And finally, a unique feature of our analysis was the availability of patient level data as this data was published and becomes available, and it becomes available as subsequent trials get published. So we

00197

performed analyses that looked at two additional points, what are the implications of using aggregate data as opposed to patient level data since that seems to provide the efficacy, and by considering sequential evidence in the trials, using the patient level data, would we be able to reach a conclusion as to overall ICD efficacy sooner. So there were a lot of these analyses that are exploring potential efficacy that Don Berry is going to present next.

So this figure demonstrates the results of combining data from all the trials using either frequentist models which are the diamonds labeled Weibull, or the Bayesian models with the little squares labeled Weibull-Bayes. The vertical extensions give a look at the 95 percent confidence intervals, and the box with different colors corresponds to different modeling approaches used to combine those trials. Within each block I put without covariate adjustments and then the next line we used covariate adjustment.

All the results showed evidence of treatment effect on overall survival and as you

00198

can see from the results, they are very similar across all the models. And although we're not showing the results here today, the estimates from combining data in small trials has a lower uncertainty as compared to those from individual trials.

Now these initial models relied on drawing assumptions as to how we accommodate trial differences, and in one extreme end we defined data assuming that the trial is similar; next we relaxed the assumption and assumed that the trial differences were accommodated with either fixed or random effects and allows that inference across specific hazard function. However, we have allowed the effect of the prognostic variables and their interaction to be similar across all trials.

So, we actually wanted to have a more flexible model to fill out across specific effects of prognostic variables, and for this we used a Bayesian hierarchical model which you heard about this morning from Dr. Lewis. And that allows, because not all subgroups are represented in all trials, for example ischemic

00199

changes in heart failure Class IV patients, we know that equivalent models cannot be estimated using traditional frequentist methods.

So this figure demonstrates several things. The different trials are shown going up the Y axis with the overall data, combining data here at the top in the black. The X axis indicates the treatment effects, with a vertical dashed line at zero meaning there was no effect with standard treatment or ICD therapy. Pretrial there is two lines showing how different priors affect the findings. Prior two is dashed, it's more informative in predicting the uncertainty, so you know, the interval is narrower.

Looking at this figure, you could actually pose two important questions. Number one, is there evidence that the devices used in the different trials differed in terms of their efficacies? And number two, controlling for ejection fraction and ischemia in the NYHA heart failure classes, are the patients within the available trials similar?

Now notice that the results may be confounded with the trial, but considering the

00200

Bayesian hierarchical model that may affect some of the differential effects across trials, we see that the ICD efficacy varies across the different trials.

Why does this, however, instill uncertainty? This could be due to differences in the devices, certainly in supplemental trials we found other trials that had more variability, but it also could be due to the patient population being different with the trial, even after controlling for the ejection fraction, ischemia and heart failure class.

To show this in another format, we show here the median hazard ratio at a 95 percent confidence interval for the effective ICD treatment on the individual trials, and then for the entire population of trials at the bottom in black. We also provide the posterior probability that the hazard ratios of mortality reduction be .8 or less, and this was considered by a panel to be a clinically important reduction in mortality.

So for example, a lower than 95 percent confidence here for the overall hazard ratio includes the value of no treatment, or

00201

includes one, with an 82 percent probability the hazard ratio is .8 or less, indicating a clinically important reduction. So we then wanted to explore whether there was evidence of the ICD with respect to the patients with different clinical characteristics. You can see the differences on this figure, and then we actually have a few of them in a row. Again on the Y axis are the different clinical trials with the combined effect at the top. The two lines again represent findings under two different priors, with a red line reflecting a more informative prior. The dot represents the median and this is the line for a 95 percent confidence interval. Things to the left of the dashed line, or the vertical dashed line, indicates that there is evidence of treatment effect, and things on the right favor control therapy. These analyses were performed using the Bayesian hierarchical model to allow further actions with differentials across models. This slide looks at the efficacy of the ICD in patients between the ages of 65 and 75. We next show the evidence of patients over

00202

75. We looked at, again, patients with ejection fraction greater than 30 percent. One thing to note is that in most of the clinical trials in their inclusion criteria, they had an ejection fraction of 35 or 40 percent as an upper bound.

We also explored the effectiveness of ICD therapy across different heart failure classes. This shows Class II, we have Class III, and then finally with Class IV. Here we actually see an example of how with a more informative prior, we're much less likely to see high absolute values for the hazard ratio. And because of the lack of patients in Class IV in the different trials, they are also a less informative prior. So the upper bounds of the intervals is valueless and they are probably too large clinically to be believable. So in order to find a more informative prior we actually narrowed down these examples, so you can see that still across these trials, there is not enough information to actually cite to the evidence of the ICD.

So this slide, again, shows a kind of ratio in the clinical trials reviewed for the

00203

Class IV patients, and note that not only is there no evidence of a significant interaction, but now there's only a 49 percent probability that the hazard ratio is .8 or less.

And finally, we evaluated the evidence for ICD effectiveness in patients with ischemic disease, and here for example you can see that the DEFINITE trial, which as I indicated before was all nonischemic patients, so I didn't actually have any ischemic patients in their trial, we're able to borrow from the other trials to substitute and actually provide an estimate, but obviously the credible interval is increased as well.

So, another feature of a Bayesian hierarchical model is that it allows for the baseline survival functions to vary from trial to trial, so this figure shows the estimated posterior baseline survival functions under each trial, and then overall trials in black.

So even controlling for ejection fraction, ischemia, age and heart failure class, the figures indicate that a patient's baseline survival differs across the different trials.

So for example, patients in the

00204

SCD-HEFT trial, shown in purple, seemed to have the best survival prognosis, while patients in the CABGPATCH and MUSTT have a poorer survival, and we found several possible explanations for this difference. The variation of the cross-trial inferences, in the type of devices, in the underlying medical care, in the patient populations, or in patients whose characteristics are currently not included in our analysis, for example, gender, hazard interval, time from MI, or a prior ventricular event.

In this slide we wanted to see if there were specific patient subgroups in which the ICD was particularly ineffective or effective. From my analysis, the evidence showed there was no evidence for differential treatment effect in the individual subgroups we looked at. So here we actually showed that there were five subgroups where the posterior possibility that the hazard ratio for mortality was less than .8 was greater than 75 percent, so it will have an effect on what your decision rule is going to be for determining effectiveness.

00205

Also note that we don't show here, but we also looked at studies that included Class IV patients, and the hazard ratio being less than .8 was actually 50 percent or less. So some of the key findings we demonstrated through these analyses is that first under all model formulations, both frequentist and Bayesian, there seemed to be evidence for the efficacy of overall survival. Second, in this particular clinical domain and intervention, evidence from Bayesian models are generally similar to those obtained under those frequentist models. Evidence obtained through combining data from all trials has lowered uncertainties compared to those from individual trials. And analyses of the combined data prove our inferences by increasing the precision of our estimates as well as the power to detect main effects and interactions. Finally, the Bayesian techniques allow us to examine questions that may not be possible under traditional frequentist methods. For example, by borrowing data across trials, we're able to examine differential effects between given

00206

patient level subgroups even if an individual trial does not include these subgroups. We next wanted to explore the Bayesian method looking at registry data. As I noted, the current ICD registry doesn't have longitudinal follow-up, so for those three methods we actually used data from the MUSTT registry, the one trial that had a registry alongside the trial, and we noted that patients in the MUSTT registry are actually both different from those in the clinical trials, as well as different characteristics from those in the ICD.

So here we'll be showing some prediction survival for patients and we're looking at different subgroups. This is for patients aged between 55 and 75, ejection fraction less than 30, New York Heart Association Class II, and with ischemic disease. And we find here both in the posterior predicted survival for the control shown in blue, and then the ICD patient shown in red, and then what we observed from these same patients in the MUSTT registry in black. And if you just focus on the control, you can

00207

see how they're reflected in the MUSTT registry.
So as you can see here, control patients in the MUSTT registry actually have better survival earlier on in the predictive prior model, but are more comparable to the predicted survival in later years. So although the Bayesian model is based on the clinical trial data, for a lot of the predicted survival experience in each of the subgroups of interest longitudinal data is so important because the clinical trial patients are often different from those in the registries. The MUSTT registry actually illustrated this point, that empirical survivor rate was quite different from what we predicted from the model.
DR. C. GOODMAN: Dr. Sanders, you had asked me for one warning.
DR. SANDERS: Okay, great.
As you attempt to borrow information across trials, the Bayesian model allows you to predict survival even if the individual trial does not include some of the subgroups, and again, this model cannot be estimated using simply frequentist methods. We just show here

00208

other subgroups with both the predicted survival and that observed in the MUSTT registry.

So finally, we turn to the analysis of the aggregate versus patient level data, and there is a lot of information in this particular figure so again, I'm going to try to orient you. On the X axis we looked at the number of trials that we said were available to combine and we assumed that the trials are combined in the order of their publication dates. We then provide estimates of the ICD effectiveness under the separate modeling assumptions. A frequentist takes the aggregate effects shown in black, and then the dashed red and blue lines refer to the Bayesian model with fixed effects, and the solid lines are Bayesian models with random effects, and the Bayesian models we did it under two different priors to allow that sensitivity offset as well.

And as you can see, with the accumulated data from trials, there is a 95 percent credible, or under both priors the posterior credibles get narrower, but the gain of information from additional data is greater

00209

than those less informative priors. And also these figures show how with two priors when combining RCT data from the trials, we can only find one line of overall ICD efficacy under one prior, but we do not rule out no efficacy under the alternative prior.

And this contrasts, this figure shows the results of analysis when taking patient level data sequentially. And as we combine data from more trials, it actually becomes more similar and precise. Using the more informative prior, we were able to see the ICD with efficacy sooner with six trials.

So something to note, while the results from aggregate Bayesian analysis are not necessarily consistent with those obtained using patient level data, their accuracy could be based on additional sources of variation, for example those that explain patient variation specifically in the study. And second, combining the data from trials sequentially, either through aggregate or patient level data, may allow us to conclude overall efficacy sooner. As already pointed out, though, and we saw it earlier today, such

00210

analyses must clarify the role of priors for reaching such a conclusion.

So, some final comments about our analyses. As we've shown, one of the main advantages of Bayesian methods is that they allow the borrowing of information across trials and subgroups, and they enable us to estimate effects within specific subgroups even if those subgroups are not represented within a given trial. Note, however, the finding is dependent on the chosen prior, and also that such analysis would not be feasible under a frequentist approach if the data in any given subgroup is not available. Also note that the availability of patient level data such as we had in our analysis allows us to directly adjust for covariates within a population, potentially explaining the differences in trial outcomes.

So, here are some of the lessons that we've learned through our analysis, and these are supported by our case study that I talked about today, but also through our literature review and the simulation studies which were performed as part of the work.

00211

First, we only want to consider claims about differential effect, subgroup effects if they're accompanied by a formal statistical test for interaction. Second, consider all sources of data in order to stipulate within the statistical model which types of interaction are likely. Third, base study design and decision-making on those subgroup effects that are likely to be strong. Fourth, if the trial-based data are sufficient, do not directly combine trial-based data with information from other sources such as observational data and/or expert opinion in a setting when you're looking for validation. When little or no trial-based information about a subgroup is available, really consider the use of other data in order to specify a prior distribution, and you will use this information to plan future studies. And finally, claims based on Bayesian methods should always include sensitivity analyses to the assumed priors. So just in summary, Bayesian approaches provide a formal method of learning

00212

from the evidence and accumulating, and we believe that incorporating these findings in the CMS decision-making processes will enable the policy makers to harness really the power of the available evidence, explore subgroup effects within a trial or across trials in a methodologically rigorous manner, assess the uncertainty of clinical trial findings, and ideally improve the health outcomes of the Medicare beneficiaries.

I will now turn it over to Don to present his findings and related analysis.

DR. BERRY: Thank you, Gillian.

DR. C. GOODMAN: Dr. Sanders, while we're waiting, Dr. Satya-Murti has a question for you, if you don't mind.

DR. SATYA-MURTI: Trying to double up here. On the interaction, can you give us a promised interaction and then if it fails, is there a way to quantify interaction, an example of what interaction you were dealing with in the strongest or the most disturbing interaction?

DR. C. GOODMAN: Can you please go to the microphone, Dr. Sanders?

00213

(Inaudible colloquy.)
DR. SATYA-MURTI: Yeah, give a clinical example so I can relate it.
DR. SANDERS: Right. So I think, I'm trying to remember a table, but the one where we showed the ones where the subgroups were greater than 75 percent probability, so I think those were ones which were actually, I think they were younger patients with low ejection fractions.
DR. SATYA-MURTI: So what was the interaction occurring?
DR. SANDERS: Oh, you mean what was the actual endpoint?
DR. SATYA-MURTI: No. Was it because they were younger and there was a third independent variable that spoiled the results?
DR. SANDERS: I'm not sure.
DR. C. GOODMAN: I'll tell you what. Why don't we hold off on answering that question. Don, are you up?
DR. BERRY: Sorry about that, take it off my time or my hide or something. These are coauthors, or this is joint with Bryan Luce, Jack Ishak and Craig Hunter of

00214

United BioSource. We were actually funded by Boston Scientific, who when they got the request from Duke for the data, said gee, what's going to happen and can you, you know, BioSource, can you use the data that's available to predict what the Duke study is going to show? And so we did that as best we could.

We used only published studies, so even though Boston Scientific has their own data, we didn't ask for that and in fact we specifically said we didn't want it, we'll do a purely literature-based analysis based on our criteria for including studies, which was randomization of ICD versus not, and all of the information that we have is publicly available.

In getting our estimates of what the survival was, we actually took out rulers and put them down on the survival curves to estimate what the values were of the various things for the individual studies. So as Scott indicates, Bayesian analysis is meta-analysis, it's inherently synthetic as you've heard, through all of the information that's available, and you do modeling. It's

00215

inherently, as Gillian said, the Bayesian fixed effects, but recognizing the uncertainty associated with the study effects and the hazards associated with periods of time is a natural thing for Bayesians.

As you see, we did a synthesis across all of the studies, we estimated the individual study effect. As Roger Lewis indicated today, this shrinkage being a, or giving rise to better estimates, and we saw some of that in Gillian's presentation, the greater precision associated with modeling that looks at results over time, and we did predictions. So here we are today at some point over the course of when these trials were approved with some information, should we do another study? And if we did with particular characteristics, what's it going to show?

So we imitated that process for each study along the way and predicted its results based on this hierarchical model. So we model the sources of variation, we look at mortality rates over time in terms of annual risks, we explore the potential time intended effect of ICD. So you'll see, it turns out that the

00216

effect is not that dependent on time. But in many cases, cancer, for example, cancer is really a heterogeneous disease, the more aggressive disease kills early, and the at risk population is therefore a more indolent form of disease and so the hazards tend to drop. So you see high hazards early on, then it drops. And in cardiovascular settings, for example I mentioned the placement of catheter in a-fib, and there's this huge recurrence of a-fib in the first month, but then the at risk population changes and it drops considerably, so we wanted to model that process. You will see that it didn't matter too much, but we incorporated it in our models.

We've accumulated data and illustrated the accrual of evidence with each study, and we answered the question, when did the evidence become conclusive, and how will we predict the next study.

So, I will come back to which studies we used. We did not know what studies Duke was going to use, we used all of the available randomized trials. The one that's not on here is MUSTT, but there are some that Gillian did

00217

not incorporate, presumably because she didn't have the data.

And I have to say that I have done meta-analysis where I have the data, for example I have the data for all the randomized trials of bone marrow transplant, both adjuvant and metastatic, and it's enormously valuable to be able to address such subgroups. So for example, young patients or some of the individual studies had shown that young patients would benefit from bone marrow transplant. If you want to know what the other study showed, it's not verified. Some had shown that HER2 negative patients might fail but other studies showed that that wasn't the case. So it's very important to have the individual patient data, and we did not have it.

Endpoint is mortality. Decomposed is not a good word for Scott to use in that case.

(Laughter.)

We dissected the Kaplan-Meier curves, did a Bayesian hierarchical model for the time of death, and we did -- this is going to be confusing to you because Gillian did several

00218

models and we too did several models. The first model -- and I'm not going to go through the formulas that you've got there. The first model is one that assumes constant hazard over the five-year period within each treatment and across the studies, except that there is a study effect that's incorporated as a covariate. It assumes the same treatment effect in all of the studies but it allows for the differential hazards over time.

The model two allows for different hazard ratios over time, so it's possible that the effect of the device, the ICD is different in the first year than in the second year, than in the third year, et cetera, and so model two allows for the possibility that the ICD effect is different in the different time periods.

Model three allows for a different effect of the treatment across the various studies. So this, you see, is the study effect and treatment effect, and this lambda stuff merely represents the different hazards, and you will see those in the pictures that I'm showing you. These are the hierarchical study

00219

effects; the one that's critical here for those of you that are into this, is this thing that Roger Lewis talked about, the hyperdistribution of the study effects, the heterogeneity of it, and the variance associated with that. That's, the conclusions are in meta-analysis very sensitive to that variable.

So model two, as I indicated, is allowing for different treatment effects over time and model three is this different study effects.

So this is model one for all of the studies, this is a relative risk of .77, so a 22 percent reduction in the risk of mortality is contemplated. The probability that ICD is effective in lowering that, there is a probability that this hazard ratio is one, is essentially one.

Model two, the time variable allowing for -- I was going to come back to this, but let me show you a picture. So this is the picture of model one versus model two. So focus on the solid lines here, so that's model one control, this is model one ICD, is solid to solid, forget about the dashes for just a

00220

moment. That's in the first year, so there was about a 17 percent mortality in the first year in the control group and about a 13 percent in the ICD group in the first year. Now these are removed from the at risk population. The second year hazard, the proportion of those who went into the second year who experienced an event in the control group, who died, was about 14 percent versus 11 percent in the device, in the ICD group in the second year. And you see that the solid line seems separated by about the same amount, and in the large odds scale it is exactly the same amount. That's model one. Model one says the benefit of the device is the same for each one of these periods. The underlying risk can differ over time, but the benefits are the same. Model two allows for, it's a completely different and independent modeling in this year than in this year and in this year, et cetera. And so it happens if you see something very similar in the first year for comparing the control versus the ICD, it's a little bit wider in the second year, you know, it's, you know, it's very similar to model one

00221

actually, except in the fourth year. In the fourth year, you know, it's a tiny bit, a tad better than the control group, and then back to, you know, the same sort of thing in year five.

So the previous slide -- oh, and this is simply the survival version, this is cut at like 50 percent of it.

So to go back, this then is the estimated relative risk and it's, again, like five different studies, combining the data from all of the trials in the five different studies. This is the relative risk in the first year, in the second year, et cetera, and that reflects the fact that there wasn't too much difference in that fourth year.

This is merely to show the study effect of model one, so we're modeling heterogeneity in the results and this, MADIT-1 had something that we tagged as being one, so this is a reference study, there is no treatment in here, this is only what is the population looking over time in these 30 studies, and what you can see is they tend to get better over time, it's not unusual.

00222

I keep fighting, I keep predicting results in breast cancer, breast cancer's getting incredibly better over time, and I'm always undershooting. This suggests the same thing.

DR. C. GOODMAN: Don, about four minutes.

DR. BERRY: Four minutes, okay. So this is a comparison allowing for the study to be different and this is just MADIT-I by itself, this is AVID by itself and what it would be, allowing for the heterogeneity in the populations, you see that the reds tend to be a little better.

This is the same page that Roger was showing you.

This is chronological risks in the model one, so it starts out, MADIT-I is the only thing that's known at that time, so this red is equal to black. The red is the Bayesian meta-analysis of the first three studies and interestingly, the effect here is about here, it's about here, you know, it's going down a little bit with time, but it's pretty predictive, so after three studies we knew more

00223

or less what the answer was going to be, and that is what that is intended to show. This is predictive analysis, so here we are with MADIT, let's predict AVID based on either model one or model three, and so model one predicts AVID to be like this thing, and this is the actual AVID. It predicts CABG to be this thing and this is the actual CABG. Coming further along you see that MADIT-II had this predicted value in model one, this predicted value in model three, and that was the answer. So the ability to do this prediction shouldn't -- and of course the widths of these things depend on the size of the trial, and so it's useful for designing trials, for instance. This is the Duke studies that were included, so they did not include CIDS or CASH or any of these, and they included this but we did not. And just to show you the comparison, this is what you saw before for model one and model two, so these numbers are exactly the same as the previous slide. This is what you get if you use the eight studies that were included in the Duke analysis but using our

00224

methodology, and the interesting thing is that the overall benefit in the Duke studies is greater, and that's partly because of MUSTT, but it's also partly because COMPANION was not included and COMPANION was not that positive, and we did include it. It still is the case that the probability of the benefit is one, and in each one of these relative risks, that advantages by ICD is improved.

So, high points, 22 percent reduction of the risk. In fact it's persistent, consistent, we saw it was known pretty early, accounted for changes in patient population. Only analyses of published data. We did no individual covariate modeling.

So I will stop, thanks.

DR. C. GOODMAN: Thank you very much, Don. Before we change our focused attention to the center mic, does our panel, do any of our panelists have a question on the presentations we just heard from Drs. Sanders and Berry before we proceed? Yes, Mark.

DR. HLATKY: I was intrigued by the fact that you did a completely independent analysis knowing that somebody was going to do

00225

an analysis, and I wonder if you would draw any conclusion about, in doing these models, whether it's a good idea to have independent replication from a separate team, given all the stuff that goes into modeling. Is that an important thing in public decisions like these?

DR. BERRY: I think it's a great idea. I talked earlier on about CISNET where we had seven modelers, they were using the same data but with different modelers, so we got to assess with the seven modelers, what is the variable in the modeling process, and it's substantial, there were differences in the various conclusions as to the relative benefits of screening and adjuvant therapy. Here there's a different dimension because Gillian had more data than I did, we used different studies, so it's apples and oranges in a way. But I think it's an absolutely important thing to assess the modeling ability and, you know, models, all models are wrong, and to assess, you know, the heterogeneity in that process by including at least a couple of models.

DR. SANDERS: I certainly agree. I

00226

think that one of the ways to validate a model is a situation like this where you can actually look at the assumptions you made, you know, Don was able to look at which of the trials we used and then see whether those particular models were going to yield similar results. And, you know, the finish to that thing that I said, that all models are wrong, but some are useful, so I think it's certainly a good exercise here.

DR. C. GOODMAN: Dr. Prager.

DR. PRAGER: Gillian, I was intrigued by the way you broke these things down and I assume that none of the studies that were done utilized any of your methodology when they went for approval of their device; is that right?

DR. SANDERS: No, they were all done using frequentist methods.

DR. PRAGER: Because if we looked closely at the CABG study the way you stratified it, it has negligible treatment effect in everything except patients with ejection fraction of greater than 30 percent and --

DR. SANDERS: Well, CABG actually is, the CABG trial was not a very positive trial.

00227

I mean, the individual trials of CABG, DINAMIT, I'm trying to think of which two, is it CASH, they all varied in terms of their individual trials in terms of effectiveness. But I think what the difference is is that in our analysis we're able to borrow information from all the other trials. And so with the CABG, although it gives you the estimate for the individual trial, there's actually more information on those types of patients from all the other trials, so it is not going to give you the same result as when you look at the CABG data from their publication.

DR. PRAGER: Okay. But nevertheless, now when you look at it this way it looks like there's very little efficacy there, and I mean to me, if I were a decision-maker making a decision on whether to approve this specific device, whatever was in this study, I would have to say it has negligible effect in everybody except those with an ejection fraction greater than 30 percent. And just, how can you see this feeding into the process?

DR. SANDERS: Right. I think this adds to verifying what your decision rule is

00228

and how you're actually going to use this Bayesian information. So that's where I tried to present that information about where the hazards ratio is going to be less than .8, so that would be seen, at least by the cardiologists on our project, as being a clinically significant reduction. So there were certainly subgroups where that happened. But for the individual subgroups there wasn't a subgroup that we could point to saying, you know, for that to be for this one group, and that's where you're getting this really great background.

DR. C. GOODMAN: Dr. Maisel.

DR. MAISEL: First of all, Gillian, as the others have mentioned, I found your presentation extremely interesting. I think that looking at the slides you had maybe about four or five from the end, those nice graphs where you did the different models and the frequentist and the Bayesian analysis, I'm struck by a couple of things.

The most striking theme to me is that it really matters what model and what prior probability you choose to use for your model,

00229

so the graded variability on that graph is between the various Bayesian models that you did, and so how do we know which model is the right model? And I know there's no answer to that, but my point simply is, it doesn't seem any different on some level from, you know, you can play statistical games and create a model that looks good or you can create a model that looks bad, so how do we know which one to believe?

DR. SANDERS: I think, actually, Don would be better to answer that.

DR. BERRY: So the good news with respect to that is, as you heard from Roger and Steve this morning, also Gillian, it's transparent. I mean, you know if you assume this, then you get that. And then you can go back and say do I want to assume this, is that, my prior, is that a reasonable prior for policy-making, and to compare the various priors. If it turns out that the answers still vary over the range of what you think are the reasonable trials, then you're not ready to make the decision. I mean, very qualitatively, you're not going to make a decision, and you

00230

may say we've got to fund a study to go out and address this question because we don't yet know the answer.
DR. C. GOODMAN: Thank you. Let us revisit, briefly, Dr. Satya-Murti's question.
DR. SANDERS: Right. So the things that we were actually exposed to, in our overall we were seeing about .65 hazard ratios. In those particular subgroups where it looked like there was at least some higher probabilities of the benefit, it ranged from about maybe .52 to .58, and in the groups with the Class IV patients it's about .8 up to .99. So, you know, it's not, it's not huge differences in the hazard ratios but there certainly are differences.
DR. SATYA-MURTI: Yeah, that helps. You have cautioned us to look for interaction, so as I understand interaction, it's a surprise third variable; is that a fair way of labeling that?
DR. BERRY: I think so. Interactions, there's a close relationship between subset analysis and interactions. So subset analysis you may ask, in this subset of patients, you

00231

know, the less than 30 percent, is there a different treatment effect than in the greater than 30 percent complementary subset? And a statistician usually tests that by way of interaction, in cancer we call it predictive markers. And it's very difficult -- I don't know the answer to this particular question, but it's very difficult to show interactions, and so statisticians become, as kind of a breed, very conservative with respect to this question.

The usual basis is that there is no interaction and to show it is very difficult, so it takes a lot of evidence to show an interaction and roughly speaking, you want to look for an extremely large or small posterior probability or extremely low P value, something that would be, in your word surprising, or there'd better be a biology associated with it.

Now, recognize that the human mind is wonderfully capable of making up biological explanations for any observation.

The third possibility is that you have to go through a confirmation study.

DR. SATYA-MURTI: The reason I was

00232

asking what that interaction, the third variable is, I'm just hoping, could we put to use, would that have a basis for further studies from here on, we need to watch out for that?

DR. SANDERS: It certainly might affect the design and what kind of patient population you might want to do the next study on. I mean, if this is, the coverage decision for ICD therapy is focusing really on patients in different New York Heart Association classes, various time from MI, which were identified kind of a priori from the existing clinical trials with subgroups where there wasn't as much evidence, and this is certainly supportive of the need for more efforts in this group.

DR. SATYA-MURTI: Thank you.

DR. GOODMAN: Okay. We're going to change --

DR. BERRY: Can I just add one more anecdote, because some people laughed when I said something this morning about subsets.

DR. C. GOODMAN: Yes, Dr. Berry.

DR. BERRY: So, ER positive/HER2

00233

negative breast cancer, that's more than 50 percent of the breast cancer. If you do a subset analysis breaking it into various pieces, you find that Taxol, any taxane does not benefit that group. We've seen it in thousands and thousands of patients. If you ask now somebody from Peoria how they treat those patients, they give them Taxol.

DR. C. GOODMAN: Thank you. We're now going to change our focus to scheduled public comments and so we will pause now while we turn toward the center mic. And our first and perhaps last scheduled public commenter is Dr. Bryan Luce. Dr. Luce is going to give his public comment, and I am reminded to say that speakers are asked to state whether or not they have financial involvement with manufacturers or other interests.

DR. LUCE: Thank you. Yes, my name is Bryan Luce. I have some financial involvement in the sense that I was a co-author of the paper you saw through Don Berry. I'm very involved with Bayesian methods development and have both industry and some public sponsorship to develop those methods.

00234

So let me begin my remarks by first thanking CMS and the committee for this opportunity to comment. More importantly, I wish to note that I'm impressed with CMS's interest in exploring novel analytical methods in a quest to improve efficiency and effectiveness in coverage decision-making, and I am particularly pleased that CMS is exploring Bayesian methods for its coverage decision process.

For purposes of disclosure, I do wish to note my long and firmly held belief that all decision processes including Medicare coverage decisions as well as the decision process itself are conceptually Bayesian processes, whether formalized or not. I also wish to disclose that I have founded the Bayesian Initiative for Health Economics and Outcomes Research, and more recently have founded and direct the PACE Initiative, which stands for pragmatic approaches to comparative effectiveness, and initially it's focusing specifically on the application of Bayesian methods in looking at comparative trials.

My statement today changes a little

00235

bit, not too much, but it specifically addresses the issue relative to conditional coverage expressed specifically by CMS's coverage and evidence development. As I was listening this morning, I would argue that the concepts of adaptive and predictive probabilities scream coverage (inaudible). As I see it, the CED process is conceptually and almost literally a Bayesian process. It is a learning and updating evidence for the decision-making process. For instance, typically if not always, CMS has chosen to consider a new clinical procedure or technology for a national coverage decision. It reviews the evidence often formally, for instance by our systematic review of literature, other existing reports, even expert opinion, and often a technology assessment from AHRQ. From the Bayesian perspective, CMS would now have an informative prior. After full review and consideration, should existing evidence be judged by CMS as promising but not sufficient, for example there may be inadequate evidence with respect to Medicare beneficiaries, which is something I

00236

think we see commonly here, additional evidence is requested before an NCD would be reconsidered.

As I understand it, the recent CED recommendation of pharmacogenomic-based warfarin followed this process, and so from a Bayesian perspective CMS now wishes to update its prior, or the existing evidence base. So this is a classic Bayesian problem or scenario; it absolutely is best treated analytically with Bayesian methods. In point of fact, I can't imagine the rationale of initiating a new trial de novo.

The CED-inspired Bayesian clinical trial should be designed in the following ways as far as I can see: First, a cap should be conceived in terms of marginality, which is adding evidence to the existing evidence base until it no longer, and I think this was talked about, and an informed decision can be made.

Second, optimally and to the extent technically feasible, the trial should allow the realtime evidence review and subsequent adapting to what is learned as the evidence accumulates, of course following their decision

00237

rules and termed a priori. An adaptive learning process literally rerandomizes treatment groups in search of optimizing therapy, which should assist CMS target coverage in an appropriate setting, patient population, providing characteristics and so forth.

Third, the trials should continue until CMS is just, and I would argue no more satisfied, that it can make an informed decision. By making the full use of existing evidence, employing realtime learning, adapting in order to optimize evidence development, and terminating as soon as CMS is satisfied, the CED process itself should be optimally efficient.

Finally, I note that this research process that I'm talking about or that we have been talking about, I think is highly consistent with a learning health care system concept that is being promoted by many, but certainly by those interested in a roundtable and evidence-based medicine.

Also, I would like to offer this point in respect to the questions you're going to be

00238

asked to answer. I think you need more questions, one of which is, and I would love to have gotten it in, except you would probably need OMB clearance. But the question I would put on the table that I would like to have you consider is to what extent do you think that Bayesian adaptive methods are applicable specifically to the CED-inspired trials? So I think that is everything, and I am very pleased to have this opportunity.

DR. C. GOODMAN: Thank you very much, Bryan Luce.

We did provide an opportunity for open public comments and I don't think anyone else has signed up. Thank you, Ms. Ellis.

So we can proceed to the next section, which is our questions to presenters. And again, let's pause for 30 seconds while all of our presenters from this morning would congregate basically in the front and center, the front row and close to the center aisle, if you would please.

We were only scheduled to have a 30-minute time slot here for questions to presenters and we may use more or less than

00239

that, and we would encourage the panel to do a couple of things. One is questions that will help us answer our remaining questions will be most welcome, because these are not trivial questions and we hope we can use our presenters' time and expertise toward that purpose. And second, concise questions are desirable, as are answers. So, I know that we sometimes have a tendency to throw in a lot of extra examples and other ideas, but we are looking for not just sensitivity but specificity here in our discussions.

With that said, Dr. Mock is first with a question, and when you throw out a question, if you have a particular speaker to whom you would like it to be addressed, please say so.

DR. MOCK: Thanks, Cliff. I want to address this to Dr. Normand, Dr. Berry, as well as Dr. Sanders, and the question is straightforward, coming from a nonstatistician. What is the decision point and what is the baseline rule on inclusion versus exclusion of studies when you roll them together, be it a meta-analysis or a Bayesian calculation?

DR. SANDERS: I know this was directed

00240

to Dr. Normand in general, but I just wanted to clear up a little bit about why we included some trials and did not with other trials. The COMPANION trial is a trial actually of ICD-CRC devices versus plain CRC versus optimal medical care, so we actually thought it was a different trial with a very different device so that we didn't think it made sense to include it in the group. The AMBIEN CAT trials are pretty small trials which, one of them looks at people with perhaps not transient heart failure, and so it was again seen as kind of a different question. But those two, certainly we could have included those and I think that the differences would have come out in the analyses, but I didn't actually feel that they were recognized as kind of the major RCT trials for us to go out for the patient level data. The CIDS and the DINAMIT trial, it certainly would have been great for us to include those, and we actually have an RO-1 being reviewed, I think today, and that includes the DINAMIT trial, and that was purely a case of us not getting access to the patient

00241

level data on time. And as Don showed, the DINAMIT trial was a trial that showed a negative, it didn't show a treatment effect, so certainly it would have increased our hazard ratio had we included those patients. So, those are just the specific trials and in general why we did them.

DR. NORMAND: I will try to be brief. I think the first cut is clinical, it's not statistical, so the clinicians need to look at the various trials and studies and determine whether or not it's a similar enough treatment to include. So, you can think of like all stents or certain stents, I'm going to throw them into one study. You can think of a drug where the dose is similar enough that I'm going to include it. So at the first cut, it really is a clinical decision that determines which studies are included.

Now short of that, then some statistical considerations would really relate to some extent the quality of the data, so I'll give you an extreme example that doesn't happen that much, but pretend there was completely missing, everybody was missing data, those

00242

types of things we would consider, but short of that it's mostly a clinical decision that would dictate what types of studies are included, and the statistician would then model those studies, assuming the data are measured similarly.

DR. BERRY: So, you heard me say earlier this morning that you use all of the information. It's a huge task and in this setting, you've got, or presumably you have to set something like what is the question you're addressing and does it include studies that address the same question. But there are low quality studies, there are high quality studies, you could include them all with a discounting for their quality.

The actual analysis, and this is one of the things that Roger Lewis indicated, the actual analysis is helpful in this regard because if you've got a study that's out to lunch, you know, it's bloated or they made up the study, and you do the hierarchical model, you will pinpoint that this thing is off the scale and that the focus isn't here, and they would borrow very little strength from that

00243

extreme. That said, it's awfully difficult to do the quality assessment, so usually we do what the usual meta-analysis folks do, we set bounds, only randomized controlled trials, and go from there.

DR. GOODMAN: Thank you, Dr. Perry. Dr. Goodman, Dr. Steve Goodman.

DR. S. GOODMAN: You didn't address it to me but I just wanted to add, I was involved a little bit in the MedCAC decision on MADIT-II and just to add a blog to the question that we're asking, in real time it's sometimes much more difficult to make these judgments than looking ex post facto, and in MADIT-II a critical issue was the expansion of the eligibility criteria to subjects who had not demonstrated inducible arrhythmia. So there was a large group there who did not have it through physiologic testing, and it was a biologic question as to whether these were biologically the same as those who had demonstrated inducibility, and at that time the study did not have enough inducibility testing in the control to assess that question.

00244

And there was a lot of questions that CMS was, and I think this was one of the reasons I was involved, was addressing, is was this combinable with the others. So somewhat done was a projection of what would have happened. That projection works if you don't include, if that inducibility factor in fact didn't make a difference. And we've learned a lot about inducibility and the predictability of the efficacy effects since then, and I gather, and Dr. Maisel could probably correct me, that it doesn't have anywhere close to the, if any, effect that it was thought to have at the time.

So I think the issue in real time of whether a given study, which is always defined to be somewhat different than previous studies, can be combined or predicted can be quite difficult, and it's very difficult to try to, as Dr. Normand said, it's fundamentally a biological and a clinical judgment, and estimates we learn more with that study and in some studies that allow us to see more clearly in retrospect, than what we could have seen at the time, because these are fundamental

00245

biological questions that are simultaneously being answered by those trials, and we may or may not have them settled by that time.

DR. C. GOODMAN: Thank you, Steve.

Dr. Hlatky is next.

DR. HLATKY: I'm trying to wrestle with the issue that might be, how do we use observational data and combine it with the trial data, and maybe an example would be helpfully concrete. I guess we may have this position often where trials are done in very specialized selected populations, and the question is whether they will work as well in less selective populations, and we might say let's start a registry and coverage for evidence development. I'm trying to see how the Bayesian process would work in that, especially saying, oh, by the way, you know, it doesn't work based on observations in these groups. Now that we have more of them or, you know, whether we had trials, we only had more people where it looked like it would work, and out in the real world people are willing to stretch that further.

So it's not to any specific speaker,

00246

but I'm wondering how we would use these methods to help us in this situation where we do coverage for evidence development.

DR. NORMAND: Well, I guess the way I would look at it is as follows: It's essentially a question of causal difference, Dr. Hlatky. We've got observational data, we've got selection issues, and so the question would be how can we use observational data in order to form how effective a particular treatment is. And so if we have multiple sources of data, so if we have some diverse populations where we think there's treatment heterogeneity, then using or adopting a Bayesian approach that tries to participate or separate those components of variance, it seems the most sensible way to proceed.

So it seems to me that there are two types of questions you have asked when I think about it, one is sort of the causal mechanism, the lack of randomization in an observational world combined with the focus in the real world, and then on top of that how the Bayesian methods could be used. And so the answer is that for a usual causal question, Bayesian or

00247

non-Bayesian, but I would submit that even in answering the causal question you could use it, because you will have a lot of heterogeneity and that using a Bayesian method would exploit that in a good way.

DR. C. GOODMAN: Thank you.

Dr. Lewis, do you have an answer to this question?

DR. LEWIS: Just to be very specific, the observational data may be comparative or noncomparative, and I think using the hierarchical modeling approach you can handle both of these, but the way the second level of the model is structured depends on what's available to you. So hypothetically if the observational data are comparative, using patients both with and without the treatment of interest, then the heterogeneity would be in the magnitude of the treatment effect, because you don't believe the patients in the observational study are fundamentally exchangeable or the same as the patients in the original RCT. So in the second level of the model, those treatment effects you wouldn't expect to fall right on top of each other.

00248

In an alternative case, which may be more common, the observational data are noncomparative, and in that setting the information from the observational data may give you information just on the rate of outcome in the control arm, or the comparative arm, or in the arm that includes the new device or drug, and then the second model looks different, but conceptually it's very similar.

DR. C. GOODMAN: Thank you. Dr. Grant is next.

DR. GRANT: A general comment, question to anyone. It seems to me that in general there are, from the simplest respect in design, there are three levels. One is, the individuals design the trial CED, all those kinds of things which you represent, some of you. The second level are those of us who spend quite a bit of our time, if not most of it, evaluating evidence from those kinds of trials. And then the third level are the decision-makers who ultimately decide to adopt, reject or to gather new evidence. You've also spent a fair amount of time telling us the perils and all the ills of P values and we

00249

haven't been involved there, and so how is it going to be different?
How is it going to be different now at all those levels? I mean, I can sort of see it in some respects generally, but how is everybody going to attain the skills necessary to be able to utilize a different way of thinking than they're used to, which they've already misused?
DR. C. GOODMAN: Dr. Berry.
DR. BERRY: So, I thought you were going to say as designers as opposed to decision-makers as to how the process would go, just a word about that, and why it is different. It's the transparency, it's the formal aspect, it's the decision analysis of why we do this, do we calculate the utilities, build this trial, make a decision, or ask for more data. The question of how you're going to actually do it, it's like in adaptive designs. I face a world out there of, speaking as kindly as I can, ignorance about this whole process, higher fees, state monitoring committees, and you can't do it overnight, you can't suddenly say we're going to change and everything is

00250

going to be this way. You have to get into it slowly, have some pilot projects, the Duke project being one, but then build up and incorporate the decision aspect and the things more formally. I mean, I loved what Brian had to say, but you can't do that right away. It's going to take time, it's going to take effort, it's going to take, you know, two modeling groups to see how it's going. You might even have a parallel process. If you don't trust it at all, you can do the usual uninformed stuff and have a parallel process where some Bayesian group educating you over time is doing this and saying boy, you shouldn't do that, but you do it anyway, and then you pay the price. It's a tough question.

DR. C. GOODMAN: Dr. Satya-Murti and then Dr. Prager. Sati?

DR. SATYA-MURTI: These are important questions for CMS. What it is in the current level of reimbursement and coverage, which is also something I come from, a Bayesian decision to cover or not cover may have been made on the basis of expert recommendations from you all,

00251

but sometime later, as Bayesian inherently does, you might decide, and we may not have a coverage decision actually, we have denied coverage. Or there's some evidence denying that. If we deny and you come back and we cover, everybody's happy. But the converse, you cover it for some time and then you decide oh, no, this is quite harmful, this has got fairly grave consequences in the media and industry.

And so that is where we're wondering if we could help CMS by building, putting in language where, just like most decisions would be conditional, pro tem, but we may reverse it and then say well, you haven't quite convinced us, like with a curfew, you have violated it, so you can't do it.

DR. C. GOODMAN: Dr. Normand.

DR. NORMAND: I had to get up on this one. This is how we assess evidence with the Bayesian approach, it has nothing to do with the philosophy.

DR. SATYA-MURTI: I agree.

DR. NORMAND: And I think one would argue, therefore, that my role is a policy

00252

world, policy-making, I'm in the department of healthcare policy and we make these types of decisions, and a lot of them are informed at your level. And so the real issue is wanting to see these issues based on the best evidence, and so I would lead that into what we do with the evidence. And some of the things that we talked about today, I think are better ways to quantify evidence, such as Bayes factors. So I think if it's going to happen anyhow, it's best therefore to try and adopt, this is a paradigm shift, is try to adopt a type of evidence building that is one that is actually going to provide the answer to the questions you seek, as opposed to having these other types of pieces of P values and whatnot, so I just wanted to say that.

DR. SATYA-MURTI: Actually, you're correct. I grant you, the language says something is reasonable and necessary, and it becomes all and none, it is not so anymore for where we are going now, all decisions to take one or other positive or negative action is likely to evolve into something that may become conditional. So if as a panel we agree on

00253

that, we could, we're in a position to recommend that more such decisions would be, I don't want to use the word conditional, but would be appropriate at that time subject to later thinking.

DR. C. GOODMAN: Dr. Lewis.

DR. LEWIS: The right thing is not always the easiest, and one of the advantages of the Bayesian approach is that it is inherently sequential in the sense that new information allows you to update yesterday's posterior and consider it as today's prior, to be further updated to a current posterior. So one of the opportunities for CMS in my opinion is that at the time that a Bayesian methodology or philosophy or approach is considered, one can take the advantage of that to explicitly state the intent to adopt a continual reassessment approach in which it is not the responsibility of CMS just to make a coverage decision initially, but to continually ensure that coverage decision remains the best decision to insure optimal outcomes and best use of the resources.

DR. C. GOODMAN: Thank you, Dr. Lewis.

00254

Dr. Prager is next.
DR. PRAGER: Most of all we have been talking about today related to devices rather than medications, and I think that may be relevant because part D is a different animal than the rest of coverage, although I think as we move forward that may actually change.
As clinicians we're often faced with the dilemma regarding off label use and I'll give you two for instances. One is drugs that are used for pain often are covered based on the etiology -- it may appear that the FDA requires you to study them for what caused the pain rather than what is the pain. So for instance, a posthepatic neuralgia is covered with three different drugs right now. That exact same pain, if it's caused by something else, is not covered at all, and as clinicians that's a problem.
One other example going to the device world is that neuropathic pain is covered with spinal cord stimulation, and yet in Europe the number one use of spinal cord stimulation is to treat the pain from the heart that is in angina, and yet in this country it's not

00255

covered there. There is a multitude of studies indicating in Europe, demonstrating efficacy of spinal cord stimulation for angina.
So what my question is, given the methodology that has been presented today, how would the group or any one of you see it applying to this off label use dilemma that many of the clinicians face?
DR. C. GOODMAN: Dr. Lewis is going to take a try at that.
DR. LEWIS: I will take a quick stab at it. When I think of off label use, I think of situations where clinicians are informally borrowing information. The on label use to a large extent consists of those specific diseases that can be defined and for which there are two phased trials leading to the labeling or the unlabeled use.
In a hierarchical approach where you borrow information, you can think of all those different disease entities that cause similar pain, for example neuropathic, like pain, as being likely to be similar in terms of their response because similarities are the underlying mechanism of the pain transmission,

00256

for example. Thus, a medication or a device that is known to have applications along, across the population of those diseases, is highly likely to be effective in a similar disease drawn from that population of diseases. In other words, the diseases are exchangeable at some level in the hierarchy, and this is what I referred to in my talk as circumstantial evidence of efficacy.

So if for example you had a disease that was relatively uncommon, that had a high morbidity associated with it, for which similar diseases, for which the device or drug had been shown to be effective in similar diseases, I would believe that you could determine with a high probability that the treatment would be effective in that disease without the need for independent evidence in that patient population. And I would urge CMS to consider those situations, especially if those diseases are relatively rare, because it allows well informed coverage decisions without the burden of separate high level evidence for each one independently.

DR. PRAGER: I completely agree with

00257

what you just said. The question is, how would you see the analysis actually getting integrated into the system for that.

DR. C. GOODMAN: Dr. Lewis.

DR. S. GOODMAN: Isn't that a question for CMS?

DR. C. GOODMAN: I would be interested in hearing his response.

DR. LEWIS: I believe that it would require the input of data to include the knowledge of the underlying disease mechanisms to define the population of diseases or the group of diseases for which treatments are likely to have similar but not identical effectiveness, and I think that's a clinical question. The goal is to create agreement on what that group of diseases is, then examine the evidence available within them so they can be integrated, so I think that's the way I see it being shown. But the first step, as many of the questions have been answered, is that clinical science has to define the domain of diseases that are thought to be similarly responsive.

DR. C. GOODMAN: Steve, but briefly.

00258

DR. S. GOODMAN: Yes. This was addressed partially taking the slide that I showed with the extrapolation from adults to children. So there's something between asking for a full-fledged clinical trial and doing what Roger suggested, which is just extending it into these other conditions, which is to say that we want some evidence, the prior evidence counts partially, we will decide collectively how much it counts, and you might be able to do a trial with 60 patients instead of 250.

So there are all grades of transferability of the prior evidence and that can be decided both scientifically and on a regulatory basis, to what extent you're going to allow that extrapolation. It has a lot of application here because of, you know, often we're involved in situations of extrapolating to older patients in the Medicare studies for which there haven't been a lot of studies, but there's been some. So there are all shades of gray between doing everything and doing nothing, and this is a place where I think it's a very rich area for application.

DR. GOODMAN: Thanks, Steve. Curtis,

00259

and then we're going to go to something else.
DR. MOCK: I want to go directly to one of the questions that the panel is going to be asked to vote on today and it specifically addresses an answer that I'm still looking for, and it's very clear. It has to do with you explaining to us the strength of the Bayesian methods and how those override the deficiencies that there may be in studies that we read and interpret, or you do, and studies that are formed in the future. And I guess a subset of the question is, this is where we are today, where are we going?
I've never met any of you before today but I have the impression that you all know each other, it's a very small supraspecific group, and it sounds like you all are believing this concept. And I think that's tremendous, but what happens five years from now or ten years from now when there's not six of you, but there's 6,000 of you trying to keep up with interpreting and correlating the recent data that we need to direct us where we need to go with patient outcomes?
So please tell us, the panel, what

00260

answer is going to be best when we're asked what our confidence is on that question, and how is that answer going to work for CMS and us as a population moving forward using the Bayesian method of statistics?

DR. C. GOODMAN: Allow me to be a little more specific. Imagine you've got a white board in front of you and we need, for starters, the three greatest potential strengths of a Bayesian approach for interpreting evidence. So we're looking for your top three here, and we're looking for an answer that does not require a statistician to comprehend, we're looking for an answer that can work within the Agency and is comprehensible to a congressional staffer who might ask.

DR. SANDERS: And you don't want the design of a trial, you want --

DR. C. GOODMAN: We want to start with interpretation of evidence, what are your top three? Dr. Lewis, do you want to take the first crack at that?

DR. LEWIS: A first crack, number one, transparency and yielding probability

00261

statements that are understandable and correctly understandable by physicians, policy-makers, regulators and the congressional staffer.

DR. C. GOODMAN: So transparency is your first?

DR. LEWIS: Transparency of probability statements, statements about the strength of evidence.

Number two is the ability to make explicit the methods by which we consider information from various sources of variable strength and quality so that that evidence can be updated as new information becomes available.

DR. AXELROD: That seems to run counter to some of the discussions we've heard about the fact that this has to be done by skilled people, and you need very experienced statisticians and people who understand how to do these things with a great deal of specificity, which suggests to me that the levels of transparency aren't quite so great as you make it out to be.

DR. C. GOODMAN: Thanks for offering

00262

that, Dr. Axelrod. We may concur with that point, but let's get these three on the table. So continue, Dr. Lewis, your explicit opinion regarding strength and qualities.

DR. LEWIS: And the third comment I would make is the explicit definition of the utility function that links the quantification of uncertainty with the ultimate decision that maximizes the benefit to the patient populations.

DR. C. GOODMAN: And if you could repeat that, I got the first part of it. Explicit definition --

DR. LEWIS: Explicit definition of the utility function which links the quantification of uncertainty with the selection of the optimal decision to maximize benefit to effective patient populations. And then with your permission, I would like to comment on the question from Dr. Axelrod.

DR. C. GOODMAN: Just not yet, please.

DR. LEWIS: Yes, sir.

DR. C. GOODMAN: So Dr. Lewis posited three such assertions, and Dr. Berry, do you

00263

have a fourth?
DR. BERRY: I just wanted to add that number two should include synthesis of the various information.
DR. NORMAND: Here's a number one. I think that one of the main benefits is that you will actually get a quantitative summary, you will get the probability that a particular treatment is better than a comparison treatment, a probability, that's number one.
DR. C. GOODMAN: Say it again.
DR. NORMAND: The probability of benefit, explicit benefit of treatment A versus treatment B, you do not get that from frequentists.
DR. C. GOODMAN: Probability of benefit.
DR. NORMAND: Yes, any size. Any size you want.
DR. C. GOODMAN: And all you get from a frequentist, at least as I understand it is kind of a thumbs up, thumbs down, not a how much.
DR. BERRY: Lewis's number -- that's Lewis's number one as well.

00264

DR. C. GOODMAN: So we have about three of your top reasons, correct, as I understand it. Yes, Dr. Normand?

DR. NORMAND: Just because -- so, this is a reason that's looking forward to the future, because in the future we will have much more data and different types of data to combine any clinical trial for any decision, so we will have genetic data, we'll have clinical data, we may have patient survey data. So in the future, because of the proliferation of databases and electronic health records, we will have much more diverse data sources to be combined, and it is not -- the Bayesian method gives you a way to do that.

DR. C. GOODMAN: So if I might rephrase that, the Bayesian method provides methodological opportunities which will be enhanced by the greater availability of data, and to use the technical term, our ability to crunch such data in the future.

DR. NORMAND: I would say that it provides, it's a method that will provide a mechanism to summarize the continuum and diverse and multiple data sources typically.

00265

DR. C. GOODMAN: That's very helpful, the continuum being everything that we will use in regard to the observational stuff to top shelf RCTs, for example. Thank you. Now, just allow me to pursue this. Thanks for the four or so swell reasons why this is the greatest thing in the world. Now we would like to hear three potential weaknesses of the Bayesian approach in the same context. I know that the several of you who've spoken that way don't typically go that way, but certainly you must have been exposed to this or even accused of it from time to time. Where are the greatest pitfalls?

DR. BERRY: So, I think Dr. Axelrod's comment about having to have trained statisticians, it's not the standard statistical approach, and if you take a Bayesian approach you have to first explain the Bayesian approach to everybody you're talking to. In addition, and I forget who said it, maybe it's related, when you get 6,000 people, there will be some good Bayesians and some not so good Bayesians. One of the things I'm

00266

constantly doing is telling people that that's a lousy Bayesian approach. So how do you, you know, you can't call Berry every time, so how are we going to do this? It takes training. So I think those are the two, and maybe it's only one weakness, but it's substantial, and this gets back to my point that we can't do this overnight because it takes training, it takes getting physicians and consumers so that they can understand what you're doing and to build credibility.

DR. C. GOODMAN: I want to make sure I understand your second point. Is another way of saying that that you've got a roomful of Bayesians and you're getting a roomful of model approaches?

DR. BERRY: I don't know if that's just another way of saying it, so it is a choice of model, but it is, the Bayesian approach and using a prior distribution and not understanding the biology, not understanding what the data shows, you can get lousy Bayesian approaches, and who is going to judge what is a lousy Bayesian approach?

DR. C. GOODMAN: So we'll be moving

00267

forward with better Bayesian approaches, and perhaps you're not in a good position now to make some sort of a judgment about their quality?

DR. BERRY: Correct.

DR. C. GOODMAN: Dr. Normand.

DR. NORMAND: So I would characterize the introduction of Bayesian approaches as a paradigm shift, and any paradigm shift is a problem, so I would say that. I'm up here to give you an anti-disadvantage, and I'm surprised that you weren't saying it's the prior. You're going to be attacked on the priors, I don't think it's a disadvantage, I actually think it's an advantage. And the reason why it's an advantage is you're making it precisely transparent. A frequentist using a prior, it's a point prior, but they don't say it.

DR. C. GOODMAN: Thank you, Dr. Normand. Dr. Goodman, is this one of the potential weaknesses?

DR. S. GOODMAN: Sort of.

(Laughter.)

The problem with the question is it

00268

doesn't say compared to what, and I can certainly list certain pitfalls.

DR. C. GOODMAN: It actually does. We will get to weigh them. So what we are looking for now is what is the downside, and then we will weigh them.

DR. S. GOODMAN: So the compared to what is the critical thing. There is almost nothing I can think of as a technical problem in the Bayesian realm that doesn't have an exact correlate in the frequentist realm.

I would say, just to amplify what Sharon said, that we're not used to talking about many of the things that we need to talk about. I'll just state some things that always, everybody knows. Minimum important, clinically important difference. I make the point that every study has 80 percent power, literally every study has 80 percent power for something. For what? And then the question is, so we routinely question 80 percent, you know, if it falls below 80 percent power. But if they say oh, it's to detect a 15 percent difference as opposed to a 13 percent, that's really uncommented on in virtually any review

00269

capacity. We're forced to talk about that and we're not used to talking about it, so is that a weakness of the Bayesian paradigm, that we must talk about things that we're uncomfortable talking about, we don't know how to do it yet? I will leave that up to you, but we don't quite know how to discuss in public or private forums some of the things that you must discuss when we talk about comprehensive Bayesian approaches.

DR. C. GOODMAN: And that hesitation to discuss that issue is not confined to Bayesians.

DR. S. GOODMAN: No, it's just rarely gone over in the other contexts.

DR. C. GOODMAN: Thank you, Dr. Goodman. Dr. Cox.

DR. COX: Following up just on this, I feel like we're in sort of medieval times and I'm sort of like a flat earth guy, and I'm looking at the new world brought to me that says it's really round, but I am challenged with reading literature. So here I pick up a paper and it describes a Bayesian analysis with all this trim that you've talked about. How do

00270

I tell -- I mean, I realize I'm ignorant, I'm on the flat planet now, but how do I tell it's a good Bayesian analysis? Is there a codification of terms in the Bayesian world that describes all of the ability to understand all the factors that go into this analysis?

DR. C. GOODMAN: Dr. Lewis. And this returns to number two on weaknesses.

DR. LEWIS: So, in many areas I join you in the flat earth society. The question is not whether or not the population of physicians who must assimilate information from the medical literature and apply that to their clinical practice are going to be able to judge the quality of Bayesian analyses. It's whether they will be better or worse at that than they are now in interpreting frequentist analyses. And I firmly believe that someone who teaches clinical medicine several times a week for many hours in an academic environment, that it is not realistic, given the complexity of current medical research, to expect practitioners planning primarily nonacademic clinical practice careers, to have them understand how to judge the quality of studies.

00271

I think that we have to have other safeguards for insuring the quality. The advantage of the Bayesian approach is that for those who are expert, it will allow better gatekeeping for the quality of the analyses, helping protect the integrity of the information presented, but I don't believee it's a method that will result in the average clinician having better insight into the strengths and the weaknesses than they now do.

DR. SANDERS: I agree with Dr. Lewis's comment, but I certainly think that there would be a place in this situation for like a user's guide for the clinical literature based on Bayesian methods. I mean, certainly there are lots of techniques out there that we need to be able to convey to the end user. You know, I work off decision analysis and having to convey what that black box means is another topic, but certainly I think the education of the end users would be something that should be looked at as we push forward.

DR. C. GOODMAN: Thank you.

Dr. Satya-Murti.

DR. SATYA-MURTI: Thanks for bringing

00272

MCID, minimum clinically important difference. What you're saying, Dr. Goodman, is that regardless of the type of analysis that you subjected the study to, is it ultimately important or not is the MCID criteria, as I dealt with it in other areas. So that being the case, even though the major advantage of Bayesian analysis is probability, it will tell you that it's 40 percent better than existing treatment, or 80 percent and so on. So if that's the major benefit of Bayesian technique, then it still falls on the clinician.

DR. C. GOODMAN: Dr. Normand.

DR. NORMAND: So, what you can do is the probability that the difference between the treatment and comparison group is bigger than X. So if X is bigger than the minimal clinical difference, you can actually put that in there, so I wanted to correct that.

DR. SATYA-MURTI: All right. If that were the case, it still behooves the decision-maker to then say, am I happy with a 20 percent improvement, am I happy with a reduction of seizures from 12 to six, or am I happy with the ability to read two more letters

00273

on a chart. So, it still falls finally on society's values and the decision-maker; is that correct?
DR. NORMAND: It always is.
DR. SANDERS: Yes.
DR. C. GOODMAN: Did you finish your point?
Dr. SATYA-MURTI: Yes. I wanted to make that for the record, that the decision-making hasn't changed regardless of these shifts in paradigms.
DR. C. GOODMAN: Thanks. Steve?
DR. S. GOODMAN: And this is just the second half of that sentence, which is yes, it is up to the decision-making side on that, but it is, the minimum clinically important difference is inherently a decision analytic construct, that is, it is a difference that offsets the safety issues and the tolerability and all of those. So if we have a complete full-blown analysis, you can be assisted by adding into the discussion of what's a minimally clinically important difference decision the analytical approaches such as Dr. Sanders and Lewis talked about.

00274

DR. SATYA-MURTI: And there's a lot of subjectivity in that.

DR. S. GOODMAN: Absolutely, which again, is the right people talking about the right things, so you're not talking about power, you're talking about how many seizures should be traded off against how much impairment driving is done by administering a particular therapy.

DR. C. GOODMAN: Thank you. In order, Dr. Dullum, Dr. Alvir and Dr. Hlatky. Dr. Dullum?

DR. DULLUM: One of the advantages, just so I can understand, is the probability that the treatment will be better. Could that also be a weakness with the ongoing analysis? I mean, would it be like the weatherman saying there's going to be a 60 percent chance that there's going to be no rain today but, oh, by the way, it did rain. So is this something that when we see the probability, that we can really rely on that as opposed to the frequentist approach? Because it's kind of an ongoing analysis, how much can you believe that probability and go down that road?

00275

DR. C. GOODMAN: Dr. Lewis.
DR. LEWIS: I will take a crack at it.
I think there are a couple of things that can happen. One is that new information becomes available and when incorporated into the prior analysis changes the probability statement, perhaps that the treatment is better controlled by a certain amount. Lacking new information, I think the probability statement is very believable because it incorporates the prior information and the available evidence.
The other strength of it and the other place that differences can come into play is if the prior information needs to be changed because, for example, we have a new understanding of the mechanism of disease, we now know that there's a common pathway that leads to this autoimmune disease and another disease that we failed to recognize as autoimmune. Now that we know there's a common mechanism, we have a different prior for believing the treatments would share in efficacies.
We can picture this kind of paradigm shift that would change the whole structure of

00276

the analysis. Short of that, I would see the probability of Bayesian being inherently more stable than the kinds of statements we make on points and null hypotheses.

DR. C. GOODMAN: Thank you. Dr. Alvir.

DR. ALVIR: Thank you. One of the things that struck me in an earlier presentation, and I forget which one it was, was that anybody can do a regression now, and I'm old enough to have been in the business when not everybody could do a regression. So given what I believe is this Bayesian creep, with more and more software out there, the fact that we have this Bayesian software, in five or ten years we could be up there doing a presentation and saying, you know, anybody can do Bayesian analyses now.

And again, you know, we have similar problems, and again, I think the classical or frequentist versus Bayesian is, for me it's an overblown argument. But you know, I think it's coming that there is going to be, you know, abuse of Bayesian methods in the future, and what can we do to prevent that? I know there

00277

is all that variability of projecting into the future, but could you at least give us some ideas?

DR. C. GOODMAN: Thank you. Just a few.

DR. SANDERS: Quickly, I think this returns to the peer review process, panels like this, peer review for journals. I mean, certainly like the Annals have some statistical reviewers for all of their articles, and certainly if this is the case they might bring on more of a Bayesian statistician as part of the review process. I think that, you know, regardless of the method, there's always going to be a point where we need to turn to those types of mechanisms.

DR. C. GOODMAN: Dr. Hlatky.

DR. HLATKY: I guess I'm going to follow up in a sense because I heard something about one of the advantages is transparency, and I think you're using the word transparency in a different way than I would use it. Transparency means that anybody can see it and understand it, and I think when you guys are saying transparency, you mean that you have

00278

these complicated mathematical functions that are laid out in some way that a highly trained expert can understand and is communicated explicitly, rather than implicitly.

So my question to you is, what about getting to the average doc understanding what this is and not saying, well, you know, you guys can cook up anything with this method?

DR. C. GOODMAN: Dr. Lewis.

DR. LEWIS: First, I agree with the way you're using transparency, and it is actually the way that I was hoping that I was using it as well. What I mean by transparency would be in terms of the user of the information derived from the analysis is that they understand what the analysis looked at. So for example, if there's a probability statement that says there's a 78 percent probability of survival six months greater with treatment A versus treatment B, that is a statement that most clinicians can understand. It's the kind of statement that they all hope the treatment analysis leads to, but it doesn't.

So the irony here in the current

00279

popular approach, we fool ourselves into thinking that what we are comfortable with is what we understand, but in fact our comfort is based on a comfort with a complete lack of understanding. So that's what I mean by transparency.

There is a second level of transparency that affects not the clinician reading the study but at the peer review process in reviewing the study, or for example another statistician attempting to replicate the analysis. And that is, the process of Bayesian analysis in many ways forces one to write down an assumption in a more explicit way because they are overt as opposed to covert, and that first of all makes it more reproducible, but most importantly invites an appropriate discussion of the merits of the assumption.

With a frequentist analysis, many of the assumptions are hidden, which completely avoids or obscures the scrutiny that they ought to undergo. Let me give you a concrete example. Every time you see longitudinal modeling with generalized estimation

00280

statements, how often do they tell you the covariate? Not very often. How many of the readers or the peer reviewers understand what effect that means on the stability of the estimates? It's a very small fraction, and yet it can change a qualitative positive result to a qualitative negative result.

DR. C. GOODMAN: Thank you, Dr. Lewis. I want to move on to what is the equivalent to our second question. You were very helpful in elucidating three to four strengths and three to four weaknesses. Now, for the purposes of designing studies, and what I'm imagining now is an enterprising and well informed sponsor for a new procedure, technology, some type of intervention, wanting to maybe approach CMS to try to figure out what sorts of evidence might CMS want to weigh in what might be a national coverage determination.

So what we're wondering, then, is what might be, how might, for the design study purpose, would this potentially strengthen, the Bayesian approach outweigh the potential liabilities in the design of a study? We're

00281

thinking about helping someone design a study, it doesn't have to be in consultation with CMS maybe, you could do it on your own, but with the Bayesian ups and downs, how do we come out on that, how confident are we that the strengths of the Bayesian approach would prevail specifically? Dr. Berry.

DR. BERRY: So, to give you the three top reasons, be adaptive, online learning, we talked about that. Using predictive probabilities, asking where is the study going, and doing it through the course of the trial. And the third one is using prior information, using parallel information that's coming from other sources during the course of the trial. So I would focus on those three.

And Dr. Goodman, the benefits, I mean, I can't give you weaknesses for design because it's so natural, it's so -- I mean, it is true that there are things that can happen along the way for the same reason of experts and non-experts designing these studies, but it makes so much sense to be looking at the data. You said what information would CMS need. You put that in the trial. You say what

00282

information does CMS need, let's build the study to give that, and if that means we can do it with 300 patients depending on the data, or if it means 3,000, maybe that's beyond the pale and we have to cut it at 1500 or something, but we will try.

DR. C. GOODMAN: Dr. Salive, do you have a question or inquiry specific to this?

DR. SALIVE: I guess this relates back to the comments earlier, so we do get companies that come in and say they're designing their trial and they are considering FDA's input and they want our input at the same time for doing a trial for a new innovative product. And the question revolves, my question revolves around prior information and sometimes we ask what that is and, you know, hand waving ensues. And the question is really, how crucial is that in this scenario, because sometimes we're told it will be a noninformative prior, other times we get some kind of rationale.

So, you know, we saw many of the analyses earlier that I thought suggested that the prior does affect somewhat the final results, and so what if they're wrong, how

00283

useful is a noninformative prior, or is this just too specific of a question?

DR. C. GOODMAN: Dr. Berry.

DR. BERRY: No, I don't think it is too specific. I can tell you this, I've been involved in many of the CRH considerations. The standard is that we use one or more priors. The issue of bringing in prior information is difficult. It's, what we've taken to do is that we have two priors, one prior for the design aspect using the information, recognizing that the FDA has its own prior, and their own prior may be noninformative. So we build the trial so that it's sufficient from the perspective of the prior distribution of the company, the experts that the company has employed or hired, but the goal is to show based on a noninformative prior that the device is effective to the extent that it's necessary. So it's a problem, it's not an easy thing to do, and there is a great deal of discussion with regulators, and there's room for bias. You know, somebody can bring a study that says I want to use this prior. You say, didn't you do some other studies that weren't

00284

quite so positive. So it's not an easy question.
DR. C. GOODMAN: Thank you. Steve Goodman, will you answer the question on the part of potential strengths outweighing potential liabilities?
DR. S. GOODMAN: Well, I was going to give the same answer to Marcel's.
DR. C. GOODMAN: Is it still relevant?
DR. S. GOODMAN: Yes.
DR. C. GOODMAN: Thank you.
Dr. Goodman.
DR. S. GOODMAN: In the earliest phases of development I think that noninformative priors are usually the way to go, and if they're not convincing to you, if an informative prior is not convincing to you, then you shouldn't allow it to be used. As Dr. Berry mentioned before, there is also learning that goes on during trials, and just the statement analytically, the two endpoints are related to each other, is in itself a prior even if you say you don't know before you start it how related.
Just that statement, this informs me

00285

about that. If this is high, if the response rate is high, I expect the mortality probably will be lower. Just that statement, without anything else you learn during the trial, that in itself in a sense is an implicit prior that allows you to learn during the trial. So it is not critical that the priors with regard to the main effect be informative; in the earliest stages of development we probably shouldn't make them informative unless they're entirely convincing.

But on the other hand, to not allow evidence-based priors when there is true prior evidence, for example, evidence that is relevant to children, children relevant to adults, or 50-year-olds relevant to 70-year-olds is an advantage to you, because you might not want more evidence than is, than common sense would require.

DR. C. GOODMAN: Thank you, Steve. Dr. Salive, is that satisfactory? Good. Further points? Dr. Dullum.

DR. DULLUM: But isn't that one of the advantages of a Bayesian, that if the prior is noninformative, as you go along further in the

00286

analysis, that that becomes less important as you get concrete information?

DR. S. GOODMAN: Yes, that was the point, but you have to have the initial linkage, you have to model something that allows the information to be borrowed as you go along. It has to be built into the design that allows you to say if this is high, you know, it tells me something about this or I will shift to a surrogate. You don't have to give your prior opinion about exactly what the nature of that relationship is going to be. You just have to say as I learn, I will allow myself to adapt to the trial. Without that statement, there is no basis for the application.

DR. C. GOODMAN: Thank you. Dr. Goodman, your voice carries pretty well, but do come to the mic whenever you have something that you want us to remember. Other points to be made with regard to this?

I'm still interested in nailing down more in our portion too whether the potential strengths outweigh the weaknesses for designing studies, and then interpreting them. Curtis Mock.

00287

DR. MOCK: I just wanted to clarify, I'm sorry I'm going back to this, I just can't get it. I think I heard you say, Dr. Berry, that one of the weaknesses of the Bayesian method is that there is no uniformity or reproducibility in how a particular meta-analysis is going to be performed by different statisticians; is that correct?

DR. BERRY: No. I didn't mean to say that. It is true, just as it's true for the frequentist methods, that if you have two statisticians that are doing the meta-analysis, they may use slightly different models, they may use different trials as part of the thing, so there is that aspect. But if two statisticians have the same prior distribution, the same kind of hierarchical setup, they're going to get the same answer.

DR. MOCK: I too work with residents on a regular basis and what I'm looking for is that power to say to the residents, don't use taxon with the HER positive/ER negative patients, and know that it's correct, and not have somebody say the opposite thing two hours ago and have the press pick it up and broadcast

00288

it. So I'm saying, is there a manner in which you can see that there would be rules for the Bayesian process that would give us uniformity of conclusions?

DR. BERRY: So, we design lots of studies, we send them to the FDA. They say, you send us a code, they may even rewrite the code, they want validation, and they want to ensure, if this is a registration trial, they want to ensure that the model is doing exactly what it says it's going to do, that they understand it, and we've had this kind of thing where everybody is happy and the thing runs great.

So it's not for the design perspective, it's not really an issue. It is reproducible. There is a certain amount of variability in the prior distribution that we use and exactly what the modeling is, are we going to do separate modeling as I did in the ICD example across the five years, or am I just going to combine all five years? Those are choices, but if two people make the same choice they're going to get the same answer.

DR. C. GOODMAN: Dr. Lewis, on this

00289

question?
DR. LEWIS: Yes. I think in terms of trying to protect yourself against the conclusions of the study being diversions, or stated and understood, the kinds of probability statements that a Bayesian analysis allows you to make are actually less prone to those changes in meaning that occur when you play the telephone game with clinical teaching.
There is a second issue that was partially addressed and I just want to clarify it, which is the need to have standards for the quality of the Bayesian analysis just like we have standards for the quality of clinical trial design, and standards for the communication of that quality, for example a consort diagram is a requirement for publication of an RCT. I believe that there is a need for some definitions regarding what is a quality reporting and conduct of a Bayesian analysis, and that will have to be developed over time and will help protect us against poor quality Bayesian analyses as the number of Bayesians increases.
DR. C. GOODMAN: Thank you, Dr. Lewis.

00290

Dr. Steve Goodman, on this point?
DR. S. GOODMAN: Yeah. I just want to address, which has come up from a number of you, the perceived problem of interpreting complicated models. The reasons these models are complicated is because the questions were complicated. It wasn't the models. Any answer that would be presented to a complicated question that put, any method that looked incredibly simple is probably wrong, or it doesn't capture the uncertainty properly.
So what you saw was an attempt to grapple with the true dimensions of uncertainty in what were inherently complicated questions. You're asking how do we combine observational, RCQ studies and RCT studies that might have five different sets of eligibility criteria and all these covariate measures? These are complicated questions, so we can choose to ignore the complexity and have a method that will give the same answer every time, you mentioned the uniform answer, or acknowledge that, you know, in reading the tea leaves there's some complexities here, and in fact there's a range of answers.

00291

So I think the assumption of your question that getting a single uniform answer is necessarily the ideal outcome to a complicated question may in itself be not the optimal model for trying to figure out what the right answer is. For a power like you, you have to deal with uncertainty, you have to make sure the uncertainties are represented properly, and that's what these models are doing, and they're complicated because the questions you're asking are complicated. You don't ask simple questions; you don't need us, and you don't need whole panels to answer simple questions.

DR. C. GOODMAN: Thank you, Steve Goodman. Dr. Normand, on this point?

DR. NORMAND: Yes. So, I just wanted to add on that last point is, I also want to emphasize that when someone summarizes the information used in a complicated analysis, the summary is not complicated, the summary can be in English and in three sentences, and you're not going to see those subscripts and subscripts, that's behind. But the point is that in actually reading the paper, you're

00292

going to see the probability. So it can be, the answer that you need to know is not going to be an equation with 3,000 subscripts.
DR. C. GOODMAN: Dr. Hlatky.
DR. HLATKY: I'm going to have to leave before too long so I'm going to say something that I think is important, and I basically think there is a place for this, but it's conditional on something very important, which, I'm convinced the more I read the material where the FDA in its guidance says the companies need to come to them when they're designing a trial, go over the information, get it locked in beforehand, and in that sense they will say okay, we're willing to deal with this analysis when you come to us for coverage. I would say that that's a good lesson for CMS, you know, if you want to use these methods in designing trials, encourage people to come in early in the design stages.
But the second thing, and I think a corollary to that is I think that it seems to me that if you want to use these methods, you have to have people on staff here in the Agency who are technical experts in these methods who

00293

can look at what's being presented and say well, this is good, this is not so good, push back, maybe get information that they analyze separately. I think based on the FDA guidance, that they sometimes want to do the analysis independently. I think that that would give an enormous amount of credibility to these things, which quite frankly are being driven by a lot of commercial, there's a lot of interests out there, there's a lot of money on the table with every one of these CMS decisions, and we shouldn't fool ourselves into thinking that there isn't.

So I mean, I think that it has to be bulletproof is really what it has to be. So I would say that I think these are encouraging techniques, but I do think that we need experts here in CMS to deal with them, and there has to be an interchange.

DR. C. GOODMAN: Thank you, Mark.

Dr. Grant.

DR. GRANT: I have to go too, so I will echo what Mark just said, and it is true. The same kind of misuse, misinterpretation and difficulties with traditional modes of

00294

synthesis and analysis, are just, you know, are just everywhere. I see it all the time and I don't think it's really any different there versus here. At the same time, I think that, if there's one point that wasn't made, and if it was maybe I missed it, was the issue of equipoise that in designing trials there is one reason, or one compelling reason, that you're exposing true patients to treatments that don't work, which I would think has a lot to say for using the design approach from that perspective.

I am entirely convinced that having direct positive statements associated with certainty will improve the system. Is there evidence to support that, I don't know. But, that's all.

DR. C. GOODMAN: Thank you, Dr. Grant. I want to pursue now, I think we've heard about, most of what we need to know for question two from our experts with their opinions. Now our question three has to do specifically with looking at whether CMS itself as an agency should incorporate evidence that uses Bayesian approaches in trials, as well as

00295

in technology assessments, and I think that technology assessments here means secondary syntheses, often depending upon systematic reviews and other secondary analyses.

And so having discussed with you first the question about the relative strengths and weaknesses vis-a-vis a frequentist approach, having talked about the net effect of the strengths and liabilities of Bayesian for designing studies and interpreting them, now let's turn to, if it's okay with the panel, turn to, well, what would be the advice to this Agency, to CMS which has to make practical decisions that will affect millions of Americans and needs to hold up to public scrutiny.

This is sort of a different world now. We're out of the classroom and graduate school and into the public fray here. So let's explore, if you will, the clinical trials piece first. CMS may be involved in, as was just said, may be talking with sponsors of interventions about how they might design clinical trials or other kinds of data gathering to ultimately inform a coverage

00296

decision, and how confident are you that the Agency might encourage that or look well upon that, using Bayesian approaches for clinical trials? Comments by our experts? Dr. Normand.

DR. NORMAND: So the answer is yes, you should do this, understanding that you have to have the expertise to do this.

DR. C. GOODMAN: Is it always yes, Dr. Normand?

DR. NORMAND: Yes.

DR. C. GOODMAN: Dr. Berry?

DR. BERRY: Yes. People talked about over time when you publish these results, I can understand them. And, you know, it's with some reluctance that I refer to something that is my own paper, but the thing that I showed the New England Journal of Medicine, a paper on breast cancer in women over 65, published May 14th, last month, that is wholly, completely Bayesian, and you can read it, you can understand it, we've gotten some comments, two reviewers said this is a wonderful study. You know, you didn't have to have a full study because you build in this Bayesian thing and you get the answer in the shortest time

00297

possible, and it's like a love fest. And it's understandable, and you should read it to see if you can understand it.

DR. C. GOODMAN: Dr. Lewis, and then Dr. Goodman.

DR. LEWIS: I agree the answer is yes and it's always yes. There are situations in which the net benefit of the Bayesian approach over a more traditional approach will be relatively less. Although there may be a leviathan situation in which the sponsor is so convinced of the effects of a therapy that they're willing to invest, almost squander extraordinary resources in the testing, and if the sponsor is willing to do that, they will end up with an easily interpretable answer. They could have obtained that answer sooner, at less cost and putting fewer patients at risk had they adopted the Bayesian approach. I don't know whether that's a concern of CMS.

DR. C. GOODMAN: Thank you. Steve Goodman.

DR. S. GOODMAN: I'll just flip the question around and say, imagine if the world was entirely Bayesian, we all understood the

00298

mechanics and the vocabulary, what reasons would we have to go to the current system? I can't imagine one.

DR. C. GOODMAN: Thank you. Dr. Mock.

DR. MOCK: Thank you for those answers regarding this question. Is there a difference in your answers regarding technology assessment versus study design? So the statistics are the same regardless.

DR. C. GOODMAN: Let the record reflect that the four experts were shaking their heads that there was no difference.

Let me pose a little bit of kind of like a second loaded question that is relevant to question three. Certainly coverage and evidence development is an important part of the set of tools or processes that the Agency has been using more or less over the years to learn as we go. And actually part of the name of this group, the ED in MedCAC is evidence development. As I believe Dr. Luce suggested, is a Bayesian approach inherent in coverage and evidence development or an important tool for it, or something that should be explicitly stated when innovators and other sponsors come

00299

to seek a national coverage determination? In other words, if it is an occasion on which the Agency might want to suggest coverage with evidence development, would the Agency might want to discuss the use of Bayesian approaches in that arrangement? Steve Goodman.

DR. S. GOODMAN: I think others have spoken to this and very well, and I just want to say two things about that. One is, certainly it is one of the only ways to coherently think about how to add up and accumulate the sometimes different information that you're going to get from the evidence development model after provisional, we'll say a provisional coverage decision, than the data that went before it. The data that went before it may have been an RCT. The data that comes after it, as was pointed out, may be an observational study, it may be noncomparative, and how to put those together is very very tricky, and I can't imagine doing it in other than this manner.

That said, and this addresses something that Dr. Satya-Murti said before,

00300

this is something you're acutely aware of, the incentives to do a proper study after even a provisional coverage assessment sometimes slip away, and this is something the FDA has come across as well. Once it's covered, there's often very very little incentive to get more evidence because you're in a holding pattern, and unless you have some sort of enforcement pattern, that is a time-limited approval under which it is required that more evidence be gathered of a certain type that actually will change your, will be sufficient to change your decision depending on how it comes out, it won't work.

So that's something for you to decide, that's not a statistical issue, but it's a very very complicated issue since your decisions themselves affect both the quantity and quality of the information that comes after, and you have to think very carefully of whether the registries or models that you set up actually generate the information that you need. Sometimes they don't. They're very well intentioned, but at the end of the day they don't give you the information that allows you

00301

to actually modify the decision when the information comes in.

DR. C. GOODMAN: Thank you, point well taken. Dr. Satya-Murti.

DR. SATYA-MURTI: Yeah, thanks. I'm glad you iterated that and asked us to consider that earlier this morning. I really think saying no is infinitely harder, so if at all possible, and our premise is, if in the recommendation dossier like FDA does, CMS would come out with that, or even build it into the executive level language, I don't mean a CFR, but within the Agency's language they define what is reasonable and necessary and so on. If we can carefully put in verbiage that says that any coverage is really time and evidence dependent, it could be annulled, and maybe that's too harsh a word, but this is a pro tem decision, and that's apparently what you are in de facto doing under the guidance or documented language.

DR. C. GOODMAN: Thank you. Other comments on this point across the panel? Dr. Dullum.

DR. DULLUM: Yes. I think if you can

00302

roll in the observational data with the written approval, you could actually use streamlined better therapy, maybe even find subgroups, I don't know that this is possible, subgroups that would also benefit from the treatment process and it would really help to direct care in that aspect.

DR. C. GOODMAN: Thank you, Dr. Dullum. Dr. Grant.

DR. GRANT: Yes, just along these lines, I think this is a point well taken for the CED. There needs to be significant attention to which parameters, which evidence needs to be informed, very careful specific attention, and I think otherwise, you're not going to be in the circumstances to be able to see an active interest and consideration being formed, and that really is avoidable. So the probability should be relatively low, we should know whether or not to proceed with adoption or reject it at that particular time.

DR. C. GOODMAN: Allow me to, with the permission of the panel, ask sort of a follow-up question for number three with regard to the advisability of employing Bayesian

00303

methods compliant with the Agency, and it has to do with innovation, and I think you touched on it earlier today, so here's the issue. The MedCAC in the past, at least some of the ones in which I participated, and coverage when we look at coverage decision-making by other major payers in the U.S. and frankly in the world, one of the issues that arises is that technologies evolve over time. So the gizmo changes over time and the extra pieces, people who apply the gizmo changes over time. Certainly in the device realm, companies invent themselves around each other all the time.

So innovation is occurring in real time as we speak here, and some of the things we've heard about in discussions between innovators, regulators and payers is that if you think there's something to think about in innovation, then innovators need signals about what's going to be expected of them over time as far as evidence requirements and other hurdles, and at the same time we need to be able to recognize that the innovation isn't the ones that we think.

00304

That said, can you address the matter of the ability of Bayesian methods, if there are any, to account for or reflect innovation as it unfolds over time, and be able to drag that into more informed coverage decision-making? Dr. Normand.

DR. NORMAND: I don't know if this is going to answer your question directly or get to all of the pieces you want, but certainly there are sources of variation that will be accounted for in the Bayesian method, and these relate to both over time it would relate to, let's say with a surgical device, surgery-specific variation. It would always relate to device-specific variation. And so there are pieces of the innovation as you model the device longitudinally, and you would try to separate out those components that you realized were changing over time.

And so again, it amounts to a complex model because you're trying to separate out lots of pieces of information that would impact on the variation of the outcomes which relate to more centers using it, perhaps different skilled surgeons using it, a different patient

00305

population using it, or the device changing over time. So again, those are all components that make the model more complex, but again, it's a natural fit to a type of Bayesian model that says okay, we can try to separate those out a little bit over time.

DR. C. GOODMAN: So, that's partially helpful. What you've answered, then, is at any given time there may be variations on a technological theme, and Bayesians can kind of identify those and follow them individually?

DR. NORMAND: I'm saying that you would have at a given point in time, in theory you would have all the longitudinal information prior to that time, so all of the changes made in that device, et cetera, those things happen, and then if you wanted to look at those things right now, if we look at what happened, it's a very complex model but in theory it could be handled within a Bayesian framework. And then the second thing you're maybe asking is about future predictions of things, with this type of change made in this mechanism, if you change the device, if you do this, if you do that, what kind of impact would

00306

that have on future types of patients, and that's more of a predictive probability going into the future.

DR. C. GOODMAN: Still drawn from your progress.

DR. NORMAND: Well, it's today posterior, because you've got to believe as a Bayesian, right now I can tell you, and I'm being a little dramatic here, but you can get the best information, have the best evidence you have available right now and based on that, you write and choose what you're thinking, that becomes your prior for future events, but that's basically what Bayesians do.

DR. C. GOODMAN: Thank you.

Dr. Berry.

DR. BERRY: This is, the Bayesian person is absolutely ideal for doing that, and in fact it was one of the main reasons that CRH had this Bayesian initiative, because they would say well, devices are so different from drugs, at least back then they were in that they would change them all the time, and we want to be borrowing information from the pervious version of the device, and one can do

00307

that in a number of ways, not the least of which is the hierarchical approach. Gillian mentioned that the person that did her Bayesian analysis on hierarchical modeling was Lurdes Inoue, and she did her dissertation at Duke and her dissertation was precisely on this question. You have a device that you've changed somewhat. You'd like to have evidence in a clinical trial or some evidence, high level evidence base that it's not changed very much and that the outcome is very similar. But you don't want to run a thousand-patient trial, so you go back to the lab and you say there are measures in the lab that may be predictive, that may be related to the performance of the device, and we concede that that's true.

But we built a model as Steve Goodman suggested here, not -- what he was talking about was longitudinally, but here it's the preclinical to the clinical. So imagine four pieces where you've got preclinical on the current device, clinical on the current device, preclinical on the modification of the device, some small amount of clinical on the

00308

modification of the device, but the totality of the evidence pulls together the question of is the modification of the device doing the same thing as the previous version. So that's just one example of the kind of thing that you can do to borrow across the various levels of the technology changing.

DR. C. GOODMAN: Okay. Could you just finish this sentence for me? So, a Bayesian approach is a credible method for assessing evidence of effectiveness of an intervention even as it is evolving over time because what.

DR. BERRY: Even as the device is evolving over time because of the possibility of modeling the relationship between the previous versions of the device along the lines of what Sharon-Lise was saying, and the current version. And you may require a high level of evidence through a clinical trial; that clinical trial could be very much smaller based on the Bayesian model.

DR. GOODMAN: Thank you, that helps. Yes, Dr. Luce?

DR. LUCE: Just to build on that, and I will try to do it in a very concrete way, I

00309

can picture a setting in which there's temporal changes in evidence regarding the device could be influenced by three things. One could be secular trends of changes in the patient population underlying risks. One could be changes in the providers that are using the device, greater personal or institutional experience with the device if there's some technical expertise required for its use. And the third would be internal changes in the device, so they actually change a design feature, or a new version of it comes out. The advantage of the Bayesian approach in quantifying the estimates of the effectiveness of the device through all of those changes is that the model can appropriately and explicitly include that structure, you can estimate the effect of each of those effects on it, that's one point. The second point is you can picture a situation in which CMS, for example, may have approved coverage for a class of device and a new device comes out that in a small trial appears to be much more effective, and a decision might be to either approve the new

00310

device and disapprove the previous ones, or just approve the new device. The borrowing of information across the population of devices will help you come up with a better estimate for the true effectiveness of the new device using the information that you already have on how that class of devices performs, and that can yield a more accurate reliable decision regarding the coverage of the new device.

DR. C. GOODMAN: Thank you. And so the borrowing phenomenon is something, too, which several of you have referred to today, the borrowing of information. Thank you. Dr. Normand.

DR. NORMAND: Yes. I just want to follow up on something that the FDA permitted, and Dr. Maisel I think is familiar with this, and this was with an OPD, operating performance criterion, so that was done without having to implement a new clinical trial, but actually compared to a lesser number, and what the FDA approved was for an independent entity to analyze all of the clinical trials, and so that was used to borrow some information from some patient populations, maybe older diabetics in

00311

some groups, but they actually used that information, permitted that information to be used when a new company came in and said I want to have a new stent, I don't want to have a clinical trial because this is similar to another one already approved on the market, what number do I need to get. And having all that information together from the other clinical trials, that helped the FDA to find a very fine difference be made, basically a line in the sand to say this is what you need to move forward. So again, there is a method when you're using Bayesian methods in this manner.

DR. C. GOODMAN: Thank you very much. I don't see any questions now from our panel. Do we have any further questions, any panelist's questions that might plumb the depth of our expertise from the front row at this point? Okay.

Dr. Salive, any further questions for the experts?

DR. SALIVE: No.

DR. C. GOODMAN: I notice that about four members of the panel have left to catch planes and so only the brave remain, so we will

00312

have a discussion among ourselves in public obviously, but let's talk about each of our three main questions, and if we do need in this discussion to refer back to our experts or even our public commenter, that would be fine. Let's return, let's go back to each of the questions starting with one, let's discuss it as needed, try to answer it, and then move on to two, and then move on to three, unless there are any objections. Dr. Salive, is that okay with you?

DR. SALIVE: That's great.

DR. C. GOODMAN: Let's take question one, and I'll just sort of rephrase it, excuse me, restate it. And it's a two-part question, and this is not a voting question, it's sort of what your answers are kind of question. And so, in assessing the strength of evidence for the effectiveness of a medical intervention that incorporates Bayesian design or analysis compared to a frequentist approach, please discuss the following, and (a) is, what are the potential greatest strengths of a Bayesian approach?

And so, which of us might want to put

00313

something on the table that we might have heard or not heard so far that stands out about a Bayesian versus a frequentist approach? Dr. Satya-Murti.

DR. SATYA-MURTI: Well, foremost is, I'm not as scared of the Bayesian approach as I was until this meeting, that's an advantage for you. But anyway, adaptability strikes me as a very useful piece of information and that we can quantify the probability with that.

What does concern me is that the intensity of training and the quality of Bayesian varies, and how do I know, that's been brought up. It's not something that cannot be overcome, but that is a concern. And what is reassuring is that such ancient questions as what are clinically important to the MCID, or how much of the benefit is important to make a decision, that decision remains, so that doesn't disenfranchise the decision-maker.

DR. C. GOODMAN: So you kind of gave a thumbs up on the adaptability and a thumbs down on sort of the variability?

DR. SATYA-MURTI: Well, yeah, the complexity of the training requirement and

00314

quality, so you have to watch out. You need built-in expertise or --

DR. C. GOODMAN: Okay, complexity in training is a potential disbenefit. Other comments on this? Dr. Dullum perhaps?

DR. DULLUM: Well, I think the strength to me is the adaptability is ongoing as we do get more information, and I think that's a huge benefit that will improve the quality of care and the use of this in guiding our management of patients, so I think that's a huge benefit.

I guess I'm now not so concerned about the prior because it does, it's kind of an ongoing, it does change and if you're wrong about your prior, it seems to be of less significance once you get conflicting data, and that's all good with the strengths of it.

And in the weakness, I guess, is the complexity and still needing to have in-depth statisticians, paperers anyway, but if this makes it easier for us to understand, I think that would also be a benefit, but that seems to me to be a downside, is that the knowledge has to be so in depth.

00315

DR. C. GOODMAN: Thank you. Dr. Cox, anything about strengths or weaknesses?
DR. COX: I really don't have anything unique to say other than it does seem to me in the world that we come into, I know there's been a growth of registries, clinical registries. One of the strengths that I have learned over the past two weeks of being introduced into this in depth, and talked about today, is the ability to blend different sources of data, and this whole source of borrowing sources more intelligently and treating patients better in that regard. And the weaknesses, I think we're beating it to death, and others have said it quite well, but it's apparent that this is complex, and so we have a lot of education and I think training coming, to be able to understand this better.
DR. C. GOODMAN: Great, thanks. Dr. Maisel.
DR. MAISEL: I mean, I agree with what's been said. I think the transparency as we have been using the term with the results and the probability is a definite advantage. I

00316

think the prior knowledge that we have of the prior data is also an advantage.

I do have some concerns, unlike some of the speakers, of the need for practicing physicians to understand the data. I think it's critically important that people who take care of patients and read journal articles can understand the data that they're looking at and understand the statistical analysis that underlies it. I don't think that's a problem that can't be overcome, and I think CMS or others could work hard at putting these types of data analyses into context with a coverage decision, perhaps have a section that explains what the analysis is if a Bayesian approach was used, why this analysis was used, and whatever else may be of interest to the practicing clinician. And I think it could be worked on in journals to better explain, and not just in statistics journals but journals that practicing clinicians use to better inform them about how a Bayesian analysis might have been applied.

DR. C. GOODMAN: Great point. So you're saying as the message is disseminated,

00317

accompany it with some translation and explanation and so forth.
DR. MAISEL: Exactly. I don't think it's enough to have so-and-so wrote something down and followed up with a statistical analysis, and we are entitled to understand what that means to some degree.
DR. C. GOODMAN: Dr. Alvir.
DR. ALVIR: Just the point that if you're doing the Bayesian approach correctly, then there is all this initial investment that you have to do with just laying out all your assumptions, which doesn't preclude that from happening nor doing that correctly too, you actually should be doing all of that also, but again, just that requirement, to actually think all of these through from the beginning and putting all of those assumptions in at the front end, at the beginning, is a great advantage.
DR. C. GOODMAN: Thanks. Just to kind of recap some of the other points that we've heard, strengths anyway, again, the transparency that's come up several times, the explicitness regarding methods and synthesis

00318

with regard to strength and quality of design and findings, specific definition of utility was brought up, and then this other point with regard to as our ability to generate data, analyze it, come up with tools for mining and analyzing data, that will just play to the strengths of Bayesian from what we heard. The weaknesses, again, we heard about having to train people to do this, we talked about the multiplicity of model approaches and how there's really a spectrum of bad to good, and what we're not really good at yet is having some sort of approach to assess bad to good, and that's something that would help with regard to sifting through the methods. The paradigm shift is something that's not the fault of any particular method, but a paradigm would have to shift, and that discontinuity brings disquiet to the stakeholders. I think that the set we just recited back to you is probably not a bad summary of the strengths and weaknesses compared to the frequentist approach. Did we miss anything else that we need to state here? Dr. Satya-Murti.

00319

DR. SATYA-MURTI: Like Bill said, a small addition at the end of an article of what they see as strengths. Most of the journals only pick one or two articles, but this one has made it a point to look at strengths and weaknesses, so I think some kind of a translation of the strengths and weaknesses to a major decision would be very good. I think that's a good idea.

DR. C. GOODMAN: Okay. Good. If I'm not mistaken, I think we just answered Question 1 as part of the record, based on what we just summarized.

Question 2 is a voting question on a one to five scale, and Question 2 asks, how confident are we that the potential strengths of Bayesian approaches outweigh the potential liabilities in the design and interpretation of a published study, and let's look at the designing studies part first. How confident are we that the potential strengths of Bayesian approaches outweigh the potential liabilities basically in designing studies?

So on that set of one to five where liabilities outweigh the strengths is a one and

00320

a five is strengths outweigh the liabilities, can you hold up your cards, please. (Members voted and the votes were recorded by staff.)

DR. C. GOODMAN: Thank you. Okay. Part two is the same question but instead of about designing studies, it's interpreting study results. So how confident are we that the potential strength of Bayesian approaches outweigh the potential liabilities of interpreting study results? (Members voted and the votes were recorded by staff.)

DR. C. GOODMAN: And before I forget, just for the record, the panelists who had to leave earlier have supplied their scores.

MS. ELLIS: Correct.

DR. C. GOODMAN: Those are in the hopper, then. Question 3 now has to do with, and we just finished discussing that a little while ago, and this is a two-part once again. And the question goes, how confident are we that CMS should incorporate evidence that uses Bayesian approaches in technology assessments

00321

in, A, clinical trials, and B, technology assessments, submitted for coverage decisions? Okay. CMS is going to get the stuff, evidence, submitted for things that may need national coverage determinations. And in the role of, in the instance of clinical trials, how confident are we that CMS ought to utilize Bayesian methods for clinical trials? Dr. Satya-Murti.

DR. SATYA-MURTI: Very briefly. So there is an assumption there that CMS has been involved in the clinical trial when the study was already designed?

DR. C. GOODMAN: I would say not necessarily so. Dr. Salive, comment? I wouldn't assume that.

DR. SALIVE: I would not assume that.

DR. SATYA-MURTI: That changes the tenor.

DR. C. GOODMAN: At least for you, Dr. Satya-Murti. Yes, Dr. Dullum?

DR. DULLUM: Can I clarify it too? They have to use Bayesian statistics or analysis to even look at it, or are you talking about applying this to a study that's already

00322

been done by a frequentist?
DR. C. GOODMAN: My reading of it would be that a sponsor or other advocate might seek a coverage decision at this level, CMS, and would supply some body of evidence, and said body of evidence would include that of clinical trials, and those clinical trials might have used Bayesian approaches. So having been presented that evidence, how confident are you that CMS ought to incorporate it in making their decision.
DR. SATYA-MURTI: We don't necessarily know how those trials were designed yet?
DR. C. GOODMAN: We don't, but presumably what we know at CMS, and sometimes with help from an evidence-based practice center, will appraise the evidence presented to it.
DR. SALIVE: I don't think we need to make this more complicated than the question is. It seems like a fairly straightforward and simple question to me, should we have CMS incorporate Bayesian analyses using their judgment in asking the proper questions about whether it's used in the right way and

00323

incorporated the proper way. I think we can rely on their judgment to do it properly.
DR. C. GOODMAN: They could appraise it. Okay. Where one is CMS ought to discourage this sort of thing, and five is they ought to encourage, on a one to five scale.
(Members voted and the votes were recorded by staff.)
DR. C. GOODMAN: So we're not quite -- excuse me. Yes, the technology assessment part, and remember, there is a distinction between clinical trials and technology assessments, and we've got our cards up.
(Members voted and the votes were recorded by staff.)
DR. C. GOODMAN: Before some final comments, Dr. Salive, is there anything that you would have wished we had covered that we still might address before we have final comments and then adjourn?
DR. SALIVE: No.
DR. C. GOODMAN: Good. Before we do adjourn, and I just want to give credit in due time to the brave ones who survived today and are still on the panel.

00324

We have gone over quite a bit in detail as far as methodologies for Bayesian versus frequentist in terms of primary data gathering, secondary data gathering, what are some indications for the Agency. Since we still have the floor insofar as it applies to advice to the Agency, do you have any additional insights or advice to the Agency about incorporating Bayesian methods in its deliberations regarding evidence requirements for coverage or how the Agency might work with intervention sponsors that might approach it with regard to coverage? Dr. Maisel?

DR. MAISEL: I'm not sure what the normal practice is for CMS, but I do think it would be valuable to put in writing your approach to Bayesian analyses, and so if you're going to permit people to come in with analyses such as this, you might outline the things you're going to judge the analysis on, such as a prespecified amount of decisions before the trial is started, that the results should be well documented in a timely fashion before the trial is completed, those are some things that I think are critical in getting quality

00325

analyses.
DR. C. GOODMAN: Great comment.
Dr. Cox.
DR. COX: Just following up on that, these documents also help those of us in the community or in research who understand, it gives us the kind of over the bow explanation that you need to understand the analysis, so it helps press the paradigm.
DR. C. GOODMAN: Great, thank you.
Dr. Satya-Murti.
DR. SATYA-MURTI: There are commonalities between the Bayesian and the frequentist approach, and that is that the final outcome of the study still doesn't change, so in that guidance it will be good also to stress what is common, so we don't necessarily think of this as too adversarial.
DR. C. GOODMAN: Thank you. My final comment for you is this. This is a great time to be looking at Bayesian methods, whether you look them or not, looking at Bayesian methods and their sister approach, adaptive trial designs and others, this is an important time to look at these. There are several things

00326

going on in our current environment that are affecting current decision-making here in the U.S., whether public payers or private, internationally, others around the world, and the various things going on with a heightened interest in a bunch of abbreviations and acronyms, health technology assessment, HTA, EBM, evidence-based medicine, as we've heard here today, CED, coverage with evidence development, comparative effectiveness research, heterogeneity of treatment effects in other subpopulations, the role of personalized medicine, and the importance of tracking innovation and accounting for how it's changing effects on patient health.

So those half dozen things that I just listed all have the potential to benefit from the insights gained by things like Bayesian methods, so I applaud the Agency for looking at this very important issue. It's timely, it is very important that you are looking at it and considering it.

I think that there is an extraordinary wealth of expertise, not only represented by the experts from whom we've heard today, but in

00327

the United States and around the world, so I think it's a very good step that CMS is looking at seriously in this current environment, as we're looking more carefully at how we assess interventions for patients' health in the United States.

So hats off to the Agency. I want to thank, including those who aren't here, Dr. Satya-Murti, Dr. Axelrod, Dr. Cox, Dr. Dullum, Dr. Grant, Dr. Hlatky, Dr. Maisel, Dr. Prager, Dr. Alvir. I want to thank our guess speakers who have been very helpful, patient and informative; that would include Don Berry, Steve Goodman, Roger Lewis, Sharon-Lise Normand and Gillian Sanders. We are very grateful for the expert guidance and good logistical planning from Dr. Marcel Salive, Maria Ellis, and the rest of the staff here at CMS. With that, Dr. Salive?

DR. SALIVE: I want to thank all those same people plus Rosemarie Hakim, plus Steve Phurrough for thinking about this idea and then leaving, not necessarily in that order, but thanks to everyone, and safe travels home.

00328

DR. C. GOODMAN: We're adjourned, thank you.
(Whereupon, the meeting adjourned at 3:58 p.m.)