U.S. FOOD AND DRUG ADMINISTRATION

and

INFECTIOUS DISEASES SOCIETY OF AMERICA

+ + + + +

ISSUES IN THE DESIGN AND CONDUCT OF CLINICAL TRIALS OF ANTIBACTERIAL DRUGS IN THE TREATMENT OF COMMUNITY-ACQUIRED PNEUMONIA

+ + + + +

WORKSHOP

+ + + + +

FRIDAY,
JANUARY 18, 2008

+ + + + +

The workshop convened at 8:00 a.m. in the Kennedy Ballroom of the Crowne Plaza Hotel, 8777 Georgia Avenue, Silver Spring, Maryland.

CO-CHAIRS:

THOMAS R. FLEMING, PhD, Professor of Biostatistics, University of Washington
EDWARD COX, MD, MPH, Director, Office of Antimicrobial Products, Office of New Drugs, CDER, FDA
DAVID GILBERT, MD, Chief of Infectious Diseases and Director of Earle A. Chiles Research Institute, Providence Portland Medical Center and Professor of Medicine, Oregon Health and Science University

RAPPORTEUR: BRAD SPELLBERG, MD, FIDSA, Assistant Professor of Medicine, Geffen School of Medicine at UCLA, Division of Infectious Diseases, Harbor-UCLA Medical Center

PANELISTS:

RICHARD WUNDERINK, MD, Professor of Medicine, Northwestern University, Feinberg School of Medicine

LIONEL A. MANDELL, MD, FRCPC, FRCP(LOND), Professor of Medicine, McMaster University

DALE BRATZLER, DO, MPH, QIOSC Medical Director, Oklahoma Foundation for Medical Quality

KEITH P. KLUGMAN, MD, Department of Global Health, Rollins School of Public Health and Division of Infectious Diseases, School of Medicine, Emory University

JOHN POWERS, III, MD, FIDSA, Science Applications International Corporation in support of the Collaborative Clinical Research Branch, NIAID, NIH, and University of Maryland School of Medicine, and George Washington School of Medicine and Health Sciences

DANIEL M. MUSHER, MD, Head of Infectious Diseases, VA Medical Center, Houston, and Professor of Medicine and Professor of Molecular Virology and Microbiology, Baylor College of Medicine

HELEN BOUCHER, MD, Director, Infectious Diseases Fellowship Program & Assistant Professor of Medicine, Division of Infectious Diseases and Geographic Medicine, Tufts University New England Medical Center

PANELISTS (continued):

MARY SINGER, MD, PhD, Medical Officer, Division of Special Pathogens and Transplant Products, CDER Office of Antimicrobial Products, FDA
PAUL AMBROSE, PharmD, FIDSA, Institute for Clinical Pharmacodynamics, Ordway Research Institute
JOHN G. BARTLETT, MD, FIDSA, Chief, Division of Infectious Disease, Johns Hopkins University School of Medicine
ROBERT TEMPLE, MD, Associate Director for Medical Policy, FDA
GLENN TILLOTSON, PHD, FCCP, Executive Director, Scientific Affairs, Replidyne, Inc.

OTHER PARTICIPANTS:

ROBERT O'NEILL, MD, FDA
JOHN S. BRADLEY, MD, FIDSA
DENNIS DIXON, PhD, NIH

C O N T E N T S

P R O C E E D I N G S

DR. GILBERT: Good morning. I appreciate everybody's attendance. We look forward to a great day, and yesterday was mild to moderate community acquired pneumonia. Today is mild enough, severe enough to get into the hospital but not in to the intensive care unit, and we're pleased that Richard Wunderink, professor of medicine, Northwestern University, is here to present the second case.

Rich.

DR. WUNDERINK: Thank you. So these are my potential conflict of interest--I actually included the American Thoracic Society and Oklahoma Foundation for Medical Quality in which I used to participate because theoretically, there's some value to those that accumulated to me. I'm going to present a case, this is actually a little bit of a synthesis case, a 65-year-old female, resident of Atlanta, so you need to know what the

epidemiology is in Atlanta, presents to the emergency department in December with some classic symptoms. Purulent sputum, shortness of breath and fever of one days' duration. Her past medical history was significant for mild COPD. She's a "35 pack-year" smoker. She continues but has cut down. Those of you that do pulmonary know that this lady would probably qualify for COPD and probably have abnormal pulmonary function test. However, she only uses a PRN bronchodilator.

She did have an exacerbation last fall, she's not exactly sure when it was, and she was treated with some unknown antibiotic. She has diabetes, on an oral agent, hypertension, she was admitted once with shortness of breath and treated for congestive heart failure on that admission, and she is obese.

Her social history is that she's sedentary, works as a domestic housecleaner. She frequently babysits her four grandchildren

when they are not in day care, although she says none of them was ill recently. She has no recent travel, no pets or hobbies. They do have a well-maintained hot tub.

Immunization history. Both the patient and her husband received influenza vaccine last fall. She does not recall getting the pneumonia vaccine, which is the common story for most patients.

She doesn't know if her grandchildren have received the pneumonia vaccination, and she does know that her children struggle financially.

On exam, she was uncomfortable, she had a frequent productive cough, dyspnea and chills. Her blood pressure was fairly well-controlled for her. She was febrile to 39.2, pulse was a 100, and regular. Her respiratory rate was 24, her oxygen saturation on room air was 89 percent, she was quickly slapped on 2 liters of oxygen and her saturation was 92 percent.

The rest of her exam, she was obese, her lungs showed definite crackles over the left lower lobe only, and she had bronchial breathing with egophony there. There were a few wheezes. There were no rales consistent with congestive heart failure. She had no gallop rhythm and no pedal edema and the rest of her exam was unremarkable.

In the usual order that we get this stuff, she actually had labs come back before chest x-ray, and her CBC clearly showed an increased white count, and 85 percent polys. I'm going to start to deviate a little bit here from what you have in your handout because we can't get bands in our hospital, so using that as a criteria for study entry isn't valid for us.

And her hemoglobin was good. Her platelets were 110,000, and she had a baseline of 180,000. Previously, PT/INR and PTT were normal. What about the radiologic evaluation?

So I changed this to a little bit

more of a real live scenario, at least in our hospital. The initial wet read by the ED radiology resident said normal size heart and clear lungs.

And the response of the ED physician is the shotgun, normal chest x-ray plus hypoxia equals PE protocol chest CT scan, and in fact that was done and showed what the clinician at the bedside would have known, and that she had left lower lobe consolidation with an air bronchogram.

And this is actually not uncommon in our institution. The interesting thing is that when the staff physician came in and read the chest x-ray, we're seeing the CT scan coming up next, the chest x-ray the next morning was read as left lower lobe infiltrate, and if you were to enroll this patient in a study, if you would have required radiology interpretation, we wouldn't have put that patient in the study.

The rest of her labs subsequently

came back. The electrolytes were, showed some hypernatremia, increased chloride, anion gap was only 13, she had an elevated BUN and creatinine.

In the day of electronic medical records, we have what her baseline is very easily, and that was normal, a year ago.

Her blood sugar was 210, so out of control but not necessarily too out of control. Blood gases confirmed the hypoxia with a PO2 of 65 on 2 liters and fairly normal acid-based status.

So if you score this lady, she's a PSI 95, which would put her into class 4, predicted mortality of that group is 9.5 percent. She got points for her BUN, for the hypoxia, and for possible CHF, depending on how much you believe that history, and she gets the age points.

So she would be in the group that should be admitted to the hospital by their criteria, by CURB-65. She gets a two for the

age and for the BUN. Once again, roughly a 6 to 7 percent mortality.

If you flip it around and say does she need to go to the intensive care unit from the new IDSA/ATS guideline, she's got two or three minor criteria. Three is what we suggest you ought to consider the ICU. The ones that she definitely has are the BUN, and if you actually calculate PO2/FIO2 ratio, assuming that 2 liters is roughly 28 percent oxygen, it's at 232, which qualifies for a minor criteria, and then the question is whether that platelet drop was significant or not.

Now the management of this patient I would say is typical for our hospital. She had a peripheral IV started and fluids were initiated with the suspicion of dehydration. She got empirical ceftriaxone, 1 gram, and azithromycin, 1 gram IV. No blood cultures of sputum cultures were ordered, and there were no other diagnostic tests.

She was admitted to a general medicine bed under the care of a hospitalist. And if you look at the CMS scoring, which is what our hospital cares about, she did great.

Her first antibiotic was given at 3 hours and 33 minutes after presentation to the ED, the delay being because of the CT scan and the error in reading the radiology report.

This is the scenario that I faced recently. The ED physician refused to allow my research coordinator to discuss a research trial with the patient because if she refused, the patient would be outside the four hour window.

And we are doing physician-specific outcomes on how fast they get their antibiotics. So he didn't want a ding on his record.

The initial antibiotic treatment was consistent with guidelines. So we get a point there. Saturation was checked. We got there. Smoking status was assessed and so the

patient was given a brochure and the contact numbers for a smoking cessation nurse.

And immunization status was assessed and RN initiated an order, placed on the chart for Pneumovax on the day of discharge. So we got all five points for this patient.

So that's the kind of clinical scenario that is actually fairly common in patients, and the questions are what is the clinical trial design most appropriate to study hospitalized CAP? And these are some of the questions listed there. Which scoring system should be used to determine severity of illness at baseline?

And for hospitalized CAP, patients with which baseline scores should be included, which diagnostic tests would be most appropriate for including patients with moderate to severe bacterial pneumonia, including Legionella, and then what are the operating characteristics of these tests.

What's the most appropriate endpoint, and when should the primary endpoint be measured, and are there any specific safety considerations for this type of study? Thank you.

DR. GILBERT: Questions or comments regarding the case scenario? The objective is to get us focused on patients who are, definitely require hospitalization but do not require direct admission to the intensive care unit.

Yes, Daniel?

DR. MUSHER: First of all, it's such a pleasure to hear a case presented so beautifully. I mean, that's classical. I don't know--

DR. GILBERT: He's a ex chief resident.

DR. MUSHER: Yes, but the present chief residents and the present residents can't do it, and the faculty doesn't expect them to, so they don't.

It's like hearing a Beethoven

sonata played well.

In our hospital, it is a requirement that a blood culture be obtained before the antibiotic is given, and it's obtained with a 90 percent reliability, and a sputum sample is submitted in about 75 percent of patients. I just thought I would mention that.

I'd also like to--are we going to get a chance to comment on something like the vaccination recommendation?, because I've just studied that. I've studied the response of patients who've recovered from pneumococcal pneumonia, pneumococcal vaccine, and I am here to tell you that it is so distressingly poor, that we need to reconsider the strategy.

DR. GILBERT: That'll be our next workshop, Daniel.

DR. MUSHER: Thanks.

DR. GILBERT: That's a big tangential.

Any other comments? So we're going

to address many facets of the patient that was presented. The first is the spectrum of the microbial etiology of hospitalized patients with community-acquired pneumonia.

We're pleased that Lionel Mandell from McMaster University in Hamilton, Ontario, was kind enough to join us.

Lionel.

DR. MANDELL: Good morning. I'd like to begin by thanking the organizers for asking me to take part in this process. I found it extremely interesting and I've learned a lot. So thank you very much.

I work at McMaster University where you can declare your religion as Christian, Jewish, Muslim or evidence-based medicine. And Dave Sackett's office was next door to mine for ten years, and he'd come by in the morning and say good morning. He'd ask me what my evidence was.

These are my conflicts of interest. The title is--you know, what really is the

question that I'm being asked to answer and what exactly does the title mean? Well, the implications of the title are that the pathogens somehow play a role in selecting the patients for entry into a therapeutic trial of an experimental drug versus a control drug.

Now that's fine and the implications are one thing, but what if that turns out not to be the case? Then which factors and which variables should we be looking at?

Well, let's look at the overall picture for a moment. This is pretty typical of practice in North America, whether it's the U.S. or Canada.

If you took all patients with community-acquired pneumonia, about 80 percent are usually appropriately treated outside the hospital, and 20 percent would come into the hospital. That breaks down as 90 percent of that 20 percent, or 18 percent overall, go to a hospital ward, or, in other words, not to

the ICU. Only 2 percent of patients go to the ICU. That's why Rich has so much free time on his hands.

But this is the group that is the sickest and has the highest mortality rates, and this is the group here, that we're going to be focusing on.

Now you could ask the question, Why do we even want to know the etiology? and it's a reasonable question to ask. But there are a lot of very good answers to it.

First of all, it allows us to give specific or directed antimicrobial therapy, so we don't have to use broad spectrum or shotgun treatment.

Also, by doing this, by collecting data, it provides a database for the local physicians, to help them in treating patients. It also helps to establish care pathways in individual hospitals, or guidelines, whether on a local or a national basis.

Also by using narrower spectrum

agents as opposed to the broader spectrum shotgun approach, it reduces antibiotic selection pressure and this may lead to a reduction in resistance, and of course it's intellectually satisfying.

Now imagine that I'm not explaining this to a group of health experts or respiratory infection experts but just to an intelligent lay audience.

And basically what I would say is look, this is the overall situation. There are only three main variables. You've got the pathogens that invade the patients, and then cause pneumonia. The patients then, usually with their pathogens, go into the hospital.

So if we're trying to look at this situation sort of with the big picture, and decide how do we select the patients for entry into a study, we could obviously focus on patient-related issues, or on pathogen-related issues, or we might even consider hospital-related issues.

But hold this thought for a few moments and we'll come back to it. But for the next few minutes, I'm going to focus on the patients and the pathogens.

Now this is just getting out of this meeting for a second to the real world, you've got an older patient, let's say in his late sixties, he was a moderate smoker, and he comes in not feeling very well, and this is his x-ray.

Any guesses as to the pathogen? Okay. I thought so.

And here's another patient who gives you a story of not feeling very well for several days, cough, it's nonproductive, a big of a headache, bit of diarrhea. Any guesses as to the pathogen? Okay.

Well, now you know how the emergency doc feels, and based on that feeling, he or she has to decide on what kind of treatment to start this person on. So the thing I want to leave you with at this point

is remember the following.

Number one. We know, and there are reasonable data to back this up, that in elderly patients with serious pneumonias, the earlier you can get treatment started, generally the better they will do.

And number two. At the time that the treatment decision is made, the physician does not know, with any degree of certainty, what the pathogen is, and certainly has no idea as to what the susceptibility is.

Okay. Let's look at the patient-related issues for a minute.

Now there's no question, that if you're trying to put patients into a therapeutic trial, you've got to have some cutoffs and say, okay, we're not going to take anybody who's hospitalized with CAP, because we know sometimes you get patients in for social reasons or whatever, and that may cloud the issue. So let's just say they need to meet some sort of severity criteria, whether

it's the PSI or CURB-65.

Mike Fine gave a fabulous talk yesterday on PSI, and this is an excellent protocol. The problem is it really is not, as you know, a true measure of severity and was initially designed to choose, not the sickest patients, but who are the patients appropriate to send home.

It's also, as you know, very heavily age-weighted, and the problem with that is that that leads to a potential underestimation of serious cases, particularly in younger patients.

Now if we look at the CURB-65, this is an easier protocol, but for Group 2, which is the group that goes into the hospital, and on the wards, not the ICU, it's not really clear how you would go about selecting any of those two criteria on which to stratify patients because it's not clear which are the most significant prognostic factors.

Okay. So let's put the patient

aside for a minute and say okay, they need to meet certain severity criteria but beyond that, I'm not sure what I'm going to go with.

Let's look at the pathogens. So what are the pathogens? Well, this was a summary that Glenn Tillotson was good enough to provide me with.

Twenty-six studies from the literature of hospitalized CAP patients, 95 percent of whom went to a hospital ward. Only about 5 percent or less went to the unit.

The total number of patients, just under 10,000, and when you look at culture-positive patients, or patients in whom they found a pathogen, and they looked at certain selected target pathogens, the ones you'd expect, in only about a third of patients did they find the pathogens.

All right. Let's look at the sort of typical or classical bacterial pathogens and then the atypicals.

These bacterial pathogens, those

patients in whom a bug was found, 82 percent of the time it was one of these. The commonest was Strep pneumo, as you'd expect, followed by H flu, and then Staph aureus.

When you look at the atypicals and those in whom a pathogen was found, it was about 18 percent of the time it was an atypical, and again, as you'd expect, Mycoplasma, Chlamydophila, or Legionella. So no big surprise there.

This is an interesting paper. This is from Tony Anzuetto's group in San Antonio. It was published online in Chest in November 2007, and will be coming out soon. But what they did was, this was a retrospective cohort study in which they looked at 730 patients who had been hospitalized for CAP, and then they looked at the pathogens in those patients admitted to the ward versus the ICU and compared them.

So 585 patients went to the wards. 145 went to the unit. If you look overall at

the 730 patients, they only found a pathogen about one out of four times.

If you focus on the subgroup that went to the wards, this was in two tertiary care university hospitals, where they looked for these bugs, they only found a pathogen one in five times, 20 percent of the time.

In the ICU, it was almost double that. It was 40 percent. But you can see that it's tough to get a pathogen, even when you're trying to in those patients admitted to hospital.

When you look at the actual breakdown, those admitted to the ward and those to the ICU, the commonest pathogens in both groups were Strep pneumo and Staph aureus. In the ward, the third most common was H flu. In the ICU, it was Pseudomonas.

So again, your chances of finding a bug in a hospitalized CAP ward patient are not that good. If you do find one, Pneumococcus is the most important. Atypicals play a role

as well.

All right. Now for a minute, let's forget about any prognostic factors or stratification or anything like that, and let's just look at the randomization process. So you've got two scenarios. Each box represents your pool of eligible patients. So again, for a second now, let's just say these patients met certain severity criteria and they were entered into the study.

You can randomize at two points but it's clear you've got to get them on treatment pretty soon.

You can randomize right away to drug A versus drug B, the experimental and control drug, or you can say, well, I really want to know how it works against pathogens, so I'll get to get results back.

So for the first couple of days you put them on a common regimen, say a combination of two drugs from different classes than drugs A and B, so that you're

covering the waterfront. Then you get the result back and then randomize to experimental or control.

The problem with this approach is that based on what we know about how antibiotics work and how they work in the lung, and pulmonary infections, etcetera. There are two real issues here. Number one. You've already, assuming you're starting them on reasonable therapy, you'll have knocked the counts way, way down.

So that's not a fair test of either A or B.

Also, you're not providing any washout period here, nor would you want to, for the drugs that you started them on. So this approach is not appropriate. So it's clear, you've got to select your patients and then randomize to drug A or B, early on.

All right. Now it's not clear that we can do this based on the patient. Okay. And it's not clear--well, we can't do

it based on the pathogen, because we don't know the pathogen, early on.

So what if we said, okay, can I do it just based on the risks for a certain pathogen or the risks for resistance, because there are pretty well-accepted risks for each of these?

Well, let's look at the risk factors for pathogens. Pneumococcus, dementia, seizure disorders, blah, blah, blah, COPD. Okay. H flu, COPD, previous antibiotics or steroids in three months. Staph aureus, underlying lung disease, previous antibiotics. Pseudomonas, pulmonary comorbidity.

You can see that there's tremendous overlap, and certainly in our hospitals, we aren't dealing with the Canadian Olympic team. We're dealing with patients who are 65 to 70, they're smokers, they've got COPD, they've had previous antibiotics. If it's a woman, for UTI. If it's a man for prostatitis, sore

throat, God knows what.

So when you look at these factors, you're thinking, okay, so I've got Mrs. Smith here in Emerg and she's at risk for--because I got all this from a lecture that I heard--Pneumococcus, H flu, Staph aureus, Legionella and Pseudomonas.

All right. A nice study by Arancibia and his group, this was in Spain, and they looked at risk factors for Gram-negative rods in CAP patients.

And if we look at the multivariate or multivariable analysis, four things pop up, and again no big surprise. Probable aspiration, previous hospitalization, previous antibiotics, and pulmonary comorbid illness.

So again, Mrs. Smith is at risk for Pneumococcus, Staph aureus, H Flu, Pseudomonas, Gram-negative rods.

When you look in the Arancibia paper, at the incidence of Gram-negative bacterial infection in CAP, based on the risk

factors, it looks, it's almost like a linear relationship with a pretty steep slope here, and when you look at the odds ratios, it's pretty clear as well.

So the more risk factors you have, the more likely you are to have Gram-negative pneumonia.

Okay. I want to enter this patient, so risk factors for pathogens probably isn't going to work too well. What about risk factors for resistance? Okay. I'm going to focus just on the pneumococcus because it's the most important pathogen and we'll look at macrolides, beta-lactams and quinolones.

Risk factors for beta-lactam resistant strep pneumo, the extremes of age--I wouldn't consider 65 an extreme anymore--but beta-lactam treatment within the last three months, exposure to a child in day care, alcoholism, medical comorbidity or immunosuppression. Pretty straightforward.

This is a very interesting paper by Vanderkooi in CID in 2005, and what he was looking at was--or the group was looking at, was the relationship of previous antibiotics in the last three months, and the relationship of the type of antibiotic and the pneumococcus and its resistant patterns if you got community-acquired pneumonia.

So the title here is "Relative risk of infection with Macrolide resistant strep pneumo based on prior antibiotic use."

So if you didn't get a prior antibiotic, then, on average, your risk of having strep pneumo that's macrolide-resistant is about 8 percent.

If you got one but it wasn't a macrolide, let's say you got a tetracycline or a cephalosporin or penicillin, it goes up very slightly.

But if you got a macrolide in the previous three months, no matter what the reason, then the likelihood is of having

macrolide-resistant strep pneumo as a cause of your pneumonia goes way up, especially if it was a long-acting macrolide such as azithromycin.

Now similarly with the fluoroquinolones, no prior antibiotic, and here again we're looking at quinolone-resistant strep pneumo. The risk is very low. If you got a prior antibiotic but it wasn't a quinolone, the risk stays pretty low.

But if you got a prior antibiotic, the prior quinolone, then the risk goes up.

Now the scale here is different. Here it's about 9 percent, whereas you'll recall in the previous slide, the macrolide, it went up to about 50 percent.

Okay. So keep in mind again, for the nonclinicians in the audience, that at the time the treatment decision is made and treatment is started, you don't know the pathogen, you don't know the susceptibilities. Okay. So I've talked about the patient, I've

talked about the pathogen, I've talked about the risks for certain pathogens and resistance.

So now the key question becomes are there data that risks for certain pathogens or resistance are prognostic factors in CAP? This is a complicated slide, coming up. No; there are no such data.

Okay. So now the key questions become how do we best select patients for entry into a therapeutic study, on what basis do we stratify, and what are the important prognostic indicators, and what are the important outcome measures which will affect prognosis and stratification?

So before I give you the answer, again, keep this in mind, that early treatment is important, especially in the older patients. We usually don't know the pathogen when we start treatment. We definitely don't know the susceptibility.

The risk factors for both patients

and resistance overlap, and there are no specific data linking risk factors or resistance to prognosis.

So this is the design that I would suggest, and this is a design that's used for all the large-scale thromboembolism studies and the large-scale cardiology studies.

You enter, you take an eligible group of patients, put them into the pool. So that decision is based on appropriate severity criteria and then you just stratify by site. Okay.

And within each site, you do a block randomization to drug A versus drug B. I think that simplifies things quite a bit, but the important thing about stratification by site is that it does a number of things. First of all, it takes into account the local epidemiology for each of the centers, but also, it balances the differences in unmeasured confounders, and these can play a major role, potentially, in how the patients

do.

So, for example, we know that if you look at different hospitals, the time to treat is going to vary significantly. We also know that the time to go from emergency to the ward will vary, and I've told a number of you, we've talked about this, in our hospital, which is pretty typical of most Canadian hospitals, it's frequently the case, if you're hospitalized with CAP, sick enough to go to the ward, you don't actually go to the ward, you stay on a stretcher in Emergency for a couple of days, and there's nothing unusual about that, and we simply refer to stretchers one, two and three as rooms one, two and three.

And of course the use of supportive measures--how well they're used, how quickly they're instituted, like fluid, oxygen, getting older patients up and getting them mobile.

So what I would suggest is that the

pathogens, at the time the treatment decision is made, really don't have major implications in terms of patient selection and that we should choose appropriate severity criteria and then stratify by site with block randomization. Thank you.

DR. GILBERT: Questions, comments for Lionel? Yes, please, Barry. Can we get his mike activated, please.

DR. EISENSTEIN: How do you deal with the need to have a window of time before that gets to be a disqualification factor?

DR. MANDELL: That's a good question, it's a very good question, in fact, and I'm not sure of the right answer. You could argue that if they've been on antibiotics for--if they just got a dose or two, then I really don't know what to do because the drug hasn't had a chance to do much.

But you could argue that if they'd been on drugs for a few days, that it's

clearly failing, if the person says I'm, you know, getting worse, blah, blah, blah. So you could say, okay, well, still enter that person. But I'm really not sure what to do because I've thought about the idea of would you stratify based on prior treatment or not. But I don't actually have a good answer, at this point, for that.

DR. GILBERT: I'm sorry, when you use the floor mikes, I'm thinking of the recording that's going on. Could you identify yourself, please.

DR. NOEL: I'm Gary Noel from Johnson & Jonson. A question about this stratification by site, and you bring up the analogy of these vascular studies where sites are enrolling 30, 40, 50 patients. It's my experience in conducting these CAP trials, that individual sites will enroll, you know, at most, a handful of patients.

So what certainty can we have, using this model, that we really are

compensating for what you think you're compensating for by site stratification?

DR. MANDELL: Again, that's an excellent question, and before I came here, we have a huge thromboembolism research group, and cardiology group, with--anyway. And I sat down with people and talked to them about this. And that's a problem, because if you do--what ideally you'd want each site, to enter a fairly significant number of patients. Now what is that number?

Nobody seems to know for sure. But maybe it should be more than 20 or 30 patients. If it's only three or four, you've got a problem. You've also got a problem--what if only you've got 30 sites across the U.S., but four of the sites, like major university centers, enter most of the patients. That reduces your generalizability.

So you're sort of caught a little bit "between a rock and a hard place." You want more sites, you want more patients in

each site to protect against the risk of just having a few, and balancing--that's why you would do block randomization. But again that's an issue to consider.

One way around that would be to--let's say you had 30 centers, five of them entered the bulk of the patients and then you've got 25 centers with, just say eight patients, on average.

You could take those 25 centers as a block, then, and analyze that, and then the other five centers. That would be one approach.

DR. MUSHER: Just to clarify, Lionel. You do believe that attempts should be made to obtain an etiologic diagnosis?

DR. MANDELL: Oh, absolutely.

DR. MUSHER: I know you do. I just wanted to bring it out. So how you start someone on the study is because of the exigencies of--

DR. MANDELL: Right. No, I'm glad

you brought that up. Thank you. Yes. I'm not being a diagnostic nihilist about this. I definitely believe that--our policy actually is that for anybody who is hospitalized with CAP, if we see them, we do get blood cultures, we do try to get a sputum, we don't waste a lot of time trying to get them to produce it, though, if they can't.

So we do try to get a pathogen, but the reality is that, right now, January 18th, 2007, these rapid diagnostic tests simply aren't available to the average physician.

DR. GILBERT: And we were just whispering up here, that we'll get into subset analysis and how that should be built into the protocols a little later in the proceedings.

So maybe we ought to go on to the next speaker.

DR. POWERS: Oh. I'm sorry.

DR. GILBERT: Yes, John?

DR. POWERS: Dr. Mandell, I want to you to clarify something at the beginning of

your slides. These two days that we're talking about are all predicated upon severity of disease, and yesterday, Dr. Fine presented, if I remember correctly, that the PSI score was initially based on evaluating baseline variables which would predict mortality, which would seem to be severity, right? the lowest one at .1 percent and the highest one was 27 percent.

DR. MANDELL: Right.

DR. POWERS: It was only secondarily, then, used to decide who gets admitted or not.

So I wanted to ask you about when you said PSI doesn't predict severity, I wasn't clear what you mean by that.

DR. MANDELL: Okay. If I did say-- I actually don't think I said that. What I said was it wasn't developed as--it's not a true severity index and it wasn't developed to pick out the sickest patients. It was actually developed to pick out the ones that

you can send home.

DR. POWERS: That was my question. I actually thought the sending home or not was actually something that was studied down the line after PSI was developed. That's not the way it was?

DR. MANDELL: No.

DR. POWERS: Okay. I mean, I remember reading this paper where they took the original 14,000 people, looked to try to predict variables and then correlated that with mortality down the line.

DR. MANDELL: Yes. The derivation protocol population was about 14,000. It was then validated in 38,000 patients. But assuming you're not Class 1, it becomes a two-step decision. But it is then to send people home.

DR. POWERS: That's what I'm not clear on. So isn't mortality a measurement of severity; right? You have higher mortality in one group than--

DR. MANDELL: No. You're right. And that's true. That's sort of the ultimate measure of severity. But still, the thing was--it's like what's your primary outcome measure? Well, their primary outcome measure was really to which ones can I send home, not who are the sickest and should appropriately go to the unit?

And, in fact, nowhere in that 1997 New England paper that Mike wrote, does it say Class 4 goes to the ward and Class 5 goes to the unit.

DR. POWERS: Right. That was my point, cause that was sort of--it was all based on morality, which there may be other measures of severity. I thought that's maybe what you were getting at. Mortality is obviously the ultimate one but maybe there's--I think George Talbot brought up yesterday, there may be some other things we're interested in as well.

DR. GILBERT: We need to move

along. I think Brad is next.

DR. SPELLBERG: Can you comment on the feasibility of placebo arm. I'm talking about feasibility of enrolling patients. We can set aside the issue of ethicality.

DR. MANDELL: Sorry. A placebo arm for these patients? No.

[Laughter]

DR. GILBERT: All right. That was quick.

George. Quickly, please.

DR. TALBOT: Yes. George Talbot. Lionel, don't go away, please. Just to go back to some of the points discussed yesterday, this dichotomy that has sort of become embedded in our terminology about mild/moderate versus severe, and how that relates to PSI, and so forth, and also the treatment effect. Where is the treatment effect large?

So could you comment on my belief, and my hypothesis, that it's pretty easy to

say who's mild. You know, most PORT one but maybe with some exceptions. And then, when you're getting into moderate and severe, you're talking about certainly five and four, and in my view probably three, and possibly some twos. So the dichotomy should be mild, especially in the context of we're talking about placebo, mild versus moderate, severe, and how that overlaps with PSI, given its limitations.

DR. MANDELL: Yes. I know, I was talking with some people last night, I can't remember exactly who, but I think I have a pretty clear idea of what "severe" is, and I usually think of severe in terms was actually put in the guidelines, that a severe CAP is somebody who has to be intubated, or is in shock and ends up in the unit.

So that's pretty clear, and I think most people who take care of patients, if you ask them, they'd say yes, that's severe. So I think I know what severe is.

And I think I know what mild is. I really don't know how to define that intermediate group, except to say it's lack of severe and lack of mild. So that's a tough group to define, and how do you explain to a medical student, or resident, or even a colleague, this is what defines the so-called moderate. It's tough.

DR. TALBOT: I'd just comment, that's probably the largest group and it's the group that's going to make it feasible, in clinical trials, to study CAP.

DR. MANDELL: Right.

DR. TALBOT: So it seems to me that we need to really define that clearly, for the purposes of clinical trials, with reliability and accuracy, and we need to have a consensus as to whether there's a large treatment effect in that group, such that we can define a Delta, and such that we can reach a conclusion about how to design our studies.

My thought, again, as a mix of drug

developer and ex-academic ideas, that, you know, that there is a treatment effect, maybe partly on mortality at one end, but also on duration of illness, and so forth, in between.

So I would think that that middle group of moderate really needs to be clear. Maybe it's by exclusion of mild and severe, is how you define it, but we need to study that.

DR. MANDELL: No, I completely agree with you and with everything you've said. I mean, the simplest way around it might be to say, okay, maybe we aren't sure of the treatment effect for the mildest ones, and based on what Mike Niederman was talking about yesterday, that if you could rule out certain patients at the extreme end who were well, and then the ones who are clearly, you know, they have to be intubated, that are in shock, that pretty well leaves you with, say, PSI 2, 3, 4, roughly.

So that might be one way to do it, or just say CURB-65, Group Two. There have

been a couple of studies published that have compared PSI and CRB-65 and CRB--CURB-65. They're pretty close. I agree with what you're saying, though.

DR. GILBERT: All right. Well, Lionel, you got us off to a good start. I think Ed's going to introduce the next speaker.

DR. COX: I'd like to invite Dale Bratzler to the podium and Dale is the QIOSC Medical Director for the Oklahoma Foundation for Medical Quality, and he's going to be talking to us today about his work with the power of the Medicare database and antibiotic selection makes a difference.

Dale.

DR. BRATZLER: All right. Good morning. Thanks for inviting me to this meeting. It's really a pleasure to be here. What I'm going to do is shift gears a little bit and talk about some of the work that we've done with large observational studies in the

Medicare population, looking at the selection of antibiotics and patient outcomes.

So let me just give you a brief background. All of the data that I'm going to share with you today is collected as a part of the Medicare Quality Improvement Organization Program.

It's a program that's built into federal law, that requires a quality improvement organization in every state to monitor the necessity and quality of care for Medicare patients. And the program has been in existence since the early 1980's.

The program gives the QIO statutory access to patient-level data. We tend, now, to focus on specific core clinical topics, so pneumonia, heart attack, heart failure, common clinical conditions in the Medicare population.

We now specifically sample at the national level, looking at the quality of care based on a number of performance measures that

are based on published guidelines, and I'm very pleased to note that many of the members of the panels today, over the past two days, have provided expert input into this project for many, many years.

The project has been in existence since 1999, and the initial data collection started in a pilot that happened in 1994.

The data collection that occurs comes into a clinical warehouse that's run by the Medicare program. Again, CMS has no access to the data, only QIOs, which have statutory federal protection around this patient-level data, and with this data I can also marry the data to all of the Medicare claims data.

So I can look at patient outcomes, rehospitalization, mortality rates. I can look at seven day, 14 day, 30 days, whatever length of time you want to look at with respect to mortality.

So let me just talk about it, and

the other point I wanted to make is that I have no financial conflicts. I do work as a contractor to the Medicare program and have no conflicts with this presentation.

So the data I'm going to share with you today is based on primarily two large sets of data that we have access to. These data were collected in 1998-1999, and 2000-2001.

There was another data set that was collected in 1994-1995, that I'll just mention briefly. This was the initial pilot project that looked at the quality of care for Medicare patients that were hospitalized.

You can see over time, just I gave, tried to give you an example of what was happening with respect to antibiotic prescription patterns.

In the Medicare population, you can see the use of beta-lactam monotherapy was progressively decreasing based on publication of new guidelines and new evidence. Use of beta-lactams and macrolides was going up,

Quinolones of course went up, their use went up dramatically, and macrolide monotherapy for hospitalized patients remains relatively uncommon.

Similarly, we saw similar trends in the ICU population. Surprisingly, we still see quite a few patients in the ICU setting, empirically treated with beta-lactam monotherapy, again macrolide monotherapy being quite uncommon in this population.

Back in 1994-95, when they initially looked at that data set, Pat Gleason and his colleagues used that data set to look at patient 30-day mortality associated with antibiotic selection. Many of you are familiar with this paper. They used third generation cephalosporins as the reference group and showed a relative reduction in 30 day mortality for patients that received a second or a third generation cephalosporin plus a macrolide or quinolone monotherapy.

So we've repeated that work, but we

have the luxury of having more risk adjustment data elements in the subsequent data sets in 1998 and 1999. The data was collected independently by a contractor to the Medicare program, a clinical data abstraction center that routinely did reabstraction for reliability of the data. All demographic data was collected on the patients. All data on risk adjustment factors. All of the risk adjustment factors from the PSI model were collected as a part of medical record review.

In addition, we captured all microbiology data and all antibiotics that were collected within the first 36 hours of hospital stay. We also collected results of sensitivity testing for any cultures that were positive.

I'm going to focus on the antibiotic piece of the work, but this is the data set that I'm primarily sharing with you today. So we did retrospective chart review of 39,000 patients in 1998 and 1999, 38,000

Medicare patients in 2000-2001. And we had a number of exclusions. If the emergency room physician didn't make a diagnosis of pneumonia when they came in, what we called the working diagnosis, we excluded the case, again because our principal focus here was on the empiric management of-- where the doctor thought the patient had pneumonia. So if there was no working diagnosis, cases were excluded.

If they came in with comfort care only, if they were being transferred from another acute care facility, again, this particular work that I'm sharing with you today focuses only on the Medicare population, 65 and older. When you look at Medicare patients below the age of 65, that's primarily patients on chronic disability of dialysis, and so we excluded that population.

Patients, if they did not have a chest x-ray consistent with pneumonia, were also excluded from the data set.

And then we had a large number of

very specific project-specific exclusions. Because now we're looking at empiric antibiotic therapy for patients who come into the hospital with pneumonia, we excluded all of the immunosuppressed patients, organ transplant, patients who were on chemotherapy or immunosuppression. We excluded patients that never got an antimicrobial during the stay or in the first 36 hours, and we excluded those patients where we were unable to determine whether they got antibiotics in an appropriate timeframe or not, or if they had multiple admissions during the study period. We only looked at the first pneumonia admission.

So the data that I'm primarily going to share with you comes from this 18,000 patients in 1998-99 and 17,000 patients in 2000-2001.

This is the patient demographics, so you can see--again, remember, we limited to Medicare patients 65 years of age or older.

So you can see the age group of the patients. About 30 percent of the patients were 85 years of age or older. 53 percent of the patients were female. I'm not going to say a lot more about the long-term care population but about 20 percent of the population that we look at in the Medicare patient population admitted with a diagnosis of pneumonia come from nursing homes, and here's the racial demographics of the population that we reviewed.

Again, we captured all of these data elements based on chart review, so all of the components of the PSI model were collected, and are a part of the datasets.

Again, we were able to do PSI risk classification. Again, there are no Class 1 patients because we didn't look at any patients that were 50 years of age or younger. We only looked at 65 and older.

You can see there was a slight shift in the demographics over the two

timeframes, with Class 5 being a little less common in 2000-2001. But generally, most of the patients were in Class 3, 4, or 5. About 10 percent of the patients that we reviewed were admitted to the intensive care unit, again, a population of only 65 or older Medicare patients.

Well, let me get to the bottom line first. There are some fairly consistent findings in our work with this data set. Again, we used third generation cephalosporin monotherapy as the reference group.

This includes ceftriaxone or cefotaxime. That was our reference group. And generally we found, as others had, and as Pat Gleason had demonstrated with the '94-'95 data set, the patients that got quinolone monotherapy, or cephalosporin plus a macrolide, had a lower 30 day morality rate. Now this is 30-day mortality. Now this is 1998-1999.

I'm not going to say anything else

about the long-term care population, other than the fact that we've looked at this data every way that we can think to look at it, with every antibiotic combination we can think of, including the new 2005 ATSI/BSA guideline recommendations of triple therapy, use of vancomycin and other agents, and simply perhaps because the power cannot detect any significant difference in patient outcomes based on antibiotic selection in the long-term care facility population.

And again we have a fairly substantial group of patients, 14,000 in 1998-99, 13,000 in 2000-2001. Again found the same thing in this group, community-dwelling patients, quinolone monotherapy, cephalosporin plus macrolide associated with lower mortality.

Interestingly, this shows up several times. The patient population is quite small but macrolide monotherapy in some of these populations, particularly in the non-

ICU setting, is, on occasion, associated with lower mortality rates than the cephalosporin monotherapy group.

I'm not going to talk a whole lot about the risk adjustment but we've risk-adjusted this data in a variety of ways. We decided to use the components of the PSI score rather than the risk classification score itself, because it seemed to be a little bit better in terms of our risk adjustment models.

But I'll show you in a moment, we've stratified the data by PSI risk class in later studies, and again, all the data have been extensively risk-adjusted. We've actually looked at this data, to look at the effect of clustering within hospitals, but remember, that even though there are 18,000 patients in one year's data set, there are 4000 hospitals reporting the data.

So the number of cases per hospital tends to be relatively small, and when we do analysis based on-- to look at the effect of

clustering of care within hospitals, we find no difference in the results.

Here's the association within patient mortality, just looking at adjusted odd ratios, we found very little effect in the 1998-1999 data set. This effect of higher mortality rate with aminoglycoside, with any administration of aminoglycoside, is not clear. It's not clear to us whether or not this represents a true effect of some problem with treating patients with aminoglycosides or the fact that perhaps the risk adjustment model simply isn't good enough to throw out the fact that these are really sick patients that are being treated with an aminoglycoside.

In 2000-2001, again, we found a slight reduction in mortality, a statistically significant reduction in mortality of cephalosporin plus macrolide as compared to the reference group. Again this is in-patient mortality.

What about 30 day mortality?

Again, the effect becomes stronger for both quinolone monotherapy and cephalosporin plus macrolide as compared to the reference group.

Similarly, in 2000-2001, quinolone monotherapy and cephalosporin plus macrolide lower mortality rates about a 34 percent relative reduction in mortality for cephalosporin plus macrolides over beta-lactam monotherapy alone.

So this effect has been fairly consistent in almost every analysis that we've done.

What about stratifying the data by discharge timeframe? Perhaps atypical organisms might be more common in certain times of the year, and we did find fairly consistently, that in October-December, January-March, when we looked at those two timeframes, the association of atypical treatment with lower mortality rates seemed to have a greater effect.

We did not find quite the same

thing during July-September discharges for hospital, and, again, I could show you more. For this particular slide, I've combined the two datasets together, and again, we've analyzed it both ways. It doesn't change the results very much, whether we do it on an individual year basis or combine the two cohorts.

Here's stratified by PSI score, so if we look at two or three versus four or five, again, the effect of quinolone monotherapy and cephalosporin plus macrolide appear to be greater in the patients with Class 4 or 5 pneumonia that were admitted to the hospital. Again, this is the combined data set of all 27,000 patients.

This is initial antibiotic selection stratified by non-ICU setting and ICU setting. So we did not--so about 11 percent of the sample here--this is 14,000 patients--went into the ICU. We did not find the association between quinolone and

cephalosporin macrolide in the ICU population but we did in the non-ICU population.

Similarly, in 2000-2001, we found the same thing-- cephalosporin plus macrolide and quinolone monotherapy as compared to third generation cephalosporin resulted in lower mortality rates. Again, we did not find the same thing in the ICU population.

We've also looked at the data for other specific subgroups. Now, again, I have probably 80 pages of Excel files here, of data that we've analyzed from these enormous datasets. But we've looked at other things. Mark Metersky, one of our colleagues from the University of Connecticut, has looked at the patients who had bacteremic pneumonia, looking at the effect of atypical treatment in patients who had bacteremia. We had about 2500 patients that had positive blood cultures with pathogens.

And again showing initial concordant antibiotic therapy, as you would

expect, if you give an antibiotic consistent-- that is effective for the organism that was cultured, resulted in a lower mortality rate, about 30 day and in-hospital mortality rate, and the addition of a macrolide, seemed to reduce mortality rate. Interestingly, we did not find the same thing with quinolones in initial atypical coverage in patients who had bacteremic pneumonia.

So those are the two large observational datasets that we have available and I'm always open to ideas about other ways to analyze the data, and ways to share the data. This is a public data set. It is, because it is patient-identified, and a part of the Medicare QIO program, I cannot release the data set, but I do have analysts available, to work with me to do additional analyses.

We do do ongoing data collection, so you heard from Rich earlier, the case presentation, I found it interesting, that he

sent a study coordinator to the Emergency Department and they wouldn't let him enroll a patient in a clinical trial because of the pneumonia quality indicator.

Let me just tell everybody--please be assured, we built in a clinical trial data exclusion to all of these national performance measures, a number of years ago. So that ER doctor was simply incorrect. If they're enrolled in a clinical trial, the patient is excluded from the performance measures.

We built that in because we do not want to suppress the ability for clinicians to do clinical trials of antibiotics.

We do do ongoing data collection. The difference is now, since 2004, hospitals self-collect the data, and because hospitals self-collect the data, and it is validated by the way, a small sample of charts are selected every year for reabstraction. The datasets are enormous. We get about 800,000 pneumonia cases a year into the data set, with all

empiric antibiotic therapy.

What I don't get any more is risk-adjustment data elements, because the data burden of hospitals having to do the data collection. But we do look at all initial antibiotics for patients admitted and subsequently discharged with a diagnosis of pneumonia.

We look to see if they're consistent with current guidelines. We exclude all patients who have health care-associated pneumonia from the denominator of our performance measure. So just to give you an example again, now you can see the ongoing trends in antibiotic delivery to hospitalized patients. This, by the way, is all--this is Medicare patients. We have 50,000 patients per quarter of Medicare alone.

But you can see beta-lactam monotherapy dropped, now, to 7.5 percent in the fourth quarter of 2006. Beta-lactams plus macrolide continues to go up, quinolone

monotherapy about 30 percent of the population. Again, I remain surprised that about 8 percent of the ICU-admitted empiric therapy is still beta-lactam monotherapy in the United States.

Length of stay has stayed very stable. So when you look at the reductions in in hospital and 30 day mortality, it doesn't appear to be, particularly in-hospital mortality does not appear to be due to reduced length of stay.

This is 90,000 Medicare patients. You can see, length of stay has stayed relatively stable since 1998. 30 day readmission rates have not changed much. 30 day mortality rate is at about 11.4 percent.

So I'd also like to specifically thank the analysts that work with me, that have done all of the statistical work on these large datasets, and again, I always am welcome to take recommendations and input about other ways to analyze this data, and make sure that

it gets into the public domain. Thank you.

DR. COX: Thank you very much, Dale.

We'll take questions for Dale, and I'll start out with one. I was struck, on a couple of your slides, with some of the community-dwelling patients, macrolide monotherapy seemed to be doing quite well, and I guess I'm wondering, your insights on that. Is that telling us that physicians can actually identify these folks who are inclined to have better prognosis, and maybe that's part of what's going on here?

DR. BRATZLER: So I think, you know, to me, the issue of macrolide--first, it's a small number of patients, I didn't put the actual numerators, denominators. But the number of patients actually receiving macrolide monotherapy is actually very small.

So I think you're probably right, that the clinician appropriately, as Lionel said earlier, he can identify the patient with

mild pneumonia, clinically. You know, it may be hard to define it but he can, as a clinician, define the patient who looks relatively well and might do well with macrolide monotherapy.

So it may be a problem with the risk adjustment model, just as we see this consistent finding of higher mortality rate with patients who get aminoglycosides.

Is that an effect of the aminoglycoside or is that a problem with the risk adjustment model, that simply doesn't identify well enough the patients who have really bad pneumonia, that are getting an aminoglycoside?

DR. COX: Thank you.

Dr. Gilbert.

DR. GILBERT: Yes. I'm always impressed with the size of the datasets. I mean, it truly is overwhelming, almost. The pneumonia endpoint, as crisp as it is--I'm sorry. The mortality endpoint, as crisp as it

is, always raises questions because it is a Medicare population, there's a lot of comorbidity, and so forth.

So I get nervous about 30 day mortality. I mean, you can die from your MI or your pulmonary embolus, or whatever. So does the mortality, as one way to look at it-- does the mortality data hold up if you look at ten day mortality as opposed to 30 day mortality? Or is there any way to factor in or factor out the death from some other cause?

In other words, the concept of attributable mortality.

DR. BRATZLER: A great question, and that's why I tried to point out that we can analyze this data with any cut point that we would like to. I don't know that we've done 10 day mortality. But we could. We could look at 10 day, two week, seven day, whatever the panel would recommend.

We can do that. We do have that ability, because we use the Medicare and the

Social Security files to actually identify the endpoint of mortality and we can look at any different cut point.

DR. COX: Okay. And John at the microphone.

DR. REX: John Rex from AstraZeneca. The large database is always very interesting and thank you for that presentation. I have a question for clarification and then I have a question.

You say this is empiric antibiotic selection. I did not see in the list of exclusions an exclusion that said physician had some strong hint--they knew it was the pneumococcus, because that would be potentially a reason for doing some--or they knew some very specific thing.

So do we truly know this is empirical? It is not based on knowing something?

DR. BRATZLER: So in 1998-99, no, we only looked at what antibiotics were given

to the patient within the first 36 hours. That last slide, where I showed you the ongoing data set, we now exclude patients from this measure if they have a positive culture or known pathogen.

So if the clinician documents it, if they have a positive urinary antigen test or anything, those patients are excluded from the denominator, going forward.

But 1998-99, 2000-2001, we looked at all antibiotics in the first 36 hours.

DR. REX: So guided therapy is going to be buried in this and you're just not going to know.

DR. BRATZLER: It could be.

DR. REX: So then that kind a leads to my question. I'm going to put on my clinician hat. You know, I spent 15 years doing clinical medicine, and I look at the aminoglycoside outcome in this and I think about 85-year-olds, and I think about a lecture that I heard as a young faculty member

in the mid '90s from David Gilbert, the title of which was "Ten reasons why ID doctors get sued."

[Laughter]

DR. REX: And the first five, perhaps, or eight, were use of aminoglycosides. And, you know, as a practicing physician, before I would write an order for aminoglycoside, you know, I would stand, I'd have to go get a cup of coffee, I'd have to--you know, it was going to scare the daylights out of me, because David Gilbert, you know--anyway, you get the point, that with an 85-year-old, the use of an aminoglycoside is absolutely not random. I'm going to have to be "pushed to the wall" before I'm going to write that order.

And I think by the year 2000, it'd be surprising if there were very many physicians who just sort of blithely ordered a little aminogly--a little gent, she's 65, she's 85, a little gent.

So that really makes me think that there is some risk factor that you've not controlled for.

And then I've learned something today that I didn't really understand. You know, John Powers is having the same confusion. I kind of thought PORT was really a mortality score. If PORT or PSI, whatever it's called, did not have as its fundamental premise, a ranking of mortality, then you've actually, if I understand what you've done, you may not have actually pulled out a true mortality predicting thing from your multivaried analysis.

If you had an APACHE score in there, I think I would understand, cause I know how APACHE was derived. It's got its flaws, but it was strongly tied to at least one kind of mortality in one setting.

So that's actually my challenge and my question for you. I don't believe the aminoglycoside stuff is random. Ergo, I

really am suspicious that you've missed out controlling some critical mortality predicting factors.

DR. BRATZLER: And that's why I acknowledge that we are concerned about that. I will tell you, moving forward, we now try to exclude patients where they have some of those Gram-negative risk factors from the data set- Pseudomonas risk. We now exclude that population of patients also.

But I do understand that.

DR. COX: Dr. Wunderink.

DR. WUNDERINK: Just a couple of points. I put that slide in there specifically for you, Dale. But my ED docs have learned the other trick too. So if they're after four hours, they call me down to put them in a study.

[Laughter]

DR. WUNDERINK: So one comment that's pertinent to Dr. Rex's comment here. I think we're getting distracted by the PSI

score here. The importance of the score is to make sure that your groups are roughly equivalent for clinical trials. The same thing that we do routinely in the ICU for these studies, using APACHE scores, or things like that.

I don't think it should be used to say this is a patient who is appropriate for this study or not, because we know ICU patients, the range of PSI scores can actually go all the way down to one, and they have a significant mortality.

So I think we need to get away from this idea of PSI as being the way to stratify into mild, moderate and severe. I'm a very simplistic critical care physician. They're mild if they're an outpatient. They're moderate if they come into the hospital. They're severe if they're admitted to me.

But in a clinical trial, you want to make sure that your groups are roughly equivalent, and so you can look at PSI, if

you're looking at severe pneumonia, it'd probably be more pertinent to look at APACHE score.

But it's not to say--it's not grading the physician for where they put the patient.

The question I have for you Dale is on your seasonal data, where you show that there's some difference in the cephalosporin macrolide in the certain times of year. But the quinolone doesn't track the same way.

And so if this is atypical--you know, quinolone, for all of the atypicals, ought to be just as good, if not better, for some of them.

So I've never understood that part of the data either. You know, I'm almost reassured to see it consistently in all the ICU patients, there's a trend toward cephalosporin macrolide actually having a better outcome even than cephalosporin quinolone as well, and I think that's

important to note.

DR. BRATZLER: Yes. The other thing that I didn't show, Rich, is that we've actually broken down by region of the country also, and we do see some differences there by region of the country.

Mortality rates tend to be higher in the South, consistently, in our datasets, than they are in the Northeast.

DR. WUNDERINK: Is it better since I've moved?

[Laughter]

DR. BRATZLER: You know, I would say that I'm not standing up here to say that this data, this observational data proves anything.

I do think it might help, though, inform clinical trials, giving your some estimates of what the treatment effect might be or what differences you might be and might expect in designing a prospective study.

I think it is useful for that, with

respect to the statistical thought about sample size and power and all of those things.

DR. COX: Okay. Dr. Echols at the microphone.

DR. ECHOLS: Yes; thank you. Roger Echols from Replidyne. The robustness of your data is, you know, based on its size, is wonderful to see, but I remind us, that when we asked Dr. Fine yesterday about attributable mortality in his studies, he said they had looked at that and they said it was maybe 50 percent.

So if you have this variety of antibiotic use which shows no trend whatsoever, or no consistent trend in mortality outcome, and only half of your mortality perhaps is attributable to the pneumonia, how can mortality be a primary endpoint in clinical trials which are going to be confounded by all the comorbidities?

And that's, you know, it's a question to Dr. Bratzler but also, you know,

it's a question of the statisticians, and I'm just concerned that I'm seeing this, the possibility of this going to a mortality endpoint for hospitalized patients, and I just don't think that the evidence really supports that the treatment of an antibiotic, certainly without a placebo arm, is going to give you that kind of ability to show non-inferiority that is related to the treatment of the drug.

DR. BRATZLER: So I think your point, though, gets back to--and so I would agree with you, that if you're designing a clinical trial, mortality might be one endpoint that you look at, but I think there have to be other endpoints.

Just these relatively small differences in mortality would require sample sizes that would be so large in a prospective study, it's probably not feasible. So I think you're going to have to have other clinical endpoints to look at in a prospective study.

The other point that Dr. Gilbert

made was that, you know, perhaps we shouldn't be looking at 30 day rates of mortality. Maybe we want to look at something that looks more closely to the actual acute care hospitalization, that might be more linked to the actual treatment of the pneumonia.

DR. ECHOLS: In your data set, do you have the cause of death?

DR. BRATZLER: No. Death is determined from the Social Security Administration data set. So we don't--we just know they died. We don't know why.

DR. ECHOLS: Okay.

DR. COX: And John?

DR. POWERS: I just want to make a point about the measurement of severity is important, because I want to get back to something that Dr. Wunderink said.

What we're going to do this afternoon is Mary's going to go through some information about what happened in, quote, unquote, severe disease in the past. To do a

rational non-inferiority trial and use that information, we then have to define severity today in current trials, in a way that's similar to the way they defined severity in the past.

So the real question isn't how do we use PSI today or what's severe today. It's how does what we do today relate to what was done in the past and how can we relate those two treatment effects together. So I just wanted to bring that up as something we'll have to keep in mind when Mary does her presentation.

DR. COX: Dr. Talbot.

DR. TALBOT: George Talbot again. Thank you. You've seen a fairly consistent effect of quinolone alone, and cephalosporin plus macrolide. Yet you're not seeing an apparent effect when a fluoroquinolone is combined with something else, or a macrolide is combined with something else, if I remember your data correctly, from the paper and from

your presentation.

Could you comment on the biological plausibility of those observations? I'm curious about your thoughts on that.

DR. BRATZLER: So my first thought is that when you look at the total number of patients, we may not have the power to detect a significant difference, because quinolone monotherapy is relatively common, cephalosporin plus a macrolide is relatively common in our data set. Macrolide with other antibiotics is not terribly common. There's that whole group of other, that I didn't go into, but there's every combination of antibiotic you can think of in that other group, which does include small numbers of patients that got macrolides with other types, or quinolones with other types of antibiotics.

But the power to detect a difference would be relatively low there. I think that's the primary thing.

Quinolone monotherapy,

interestingly, in the bacteremic population, we did not find the association with reduced mortality as we did with cephalosporin plus macrolide.

You know, in the paper we talked about some of the reasons that that might be but we don't know what that effect is.

Dr. COX: At the microphone.

DR. DANKNER: Wayne Dankner from PAREXEL. There was a small group of patients in your exclusion criteria that the group here may be interested in, and that was about the 350 patients who did not get antibiotics within the first 36 hours. And I'm wondering if there's a possibility that that group could be better abstracted because it may provide us some information about delay versus immediate therapy, if we were thinking of a design in the future.

It'll be interesting to see what kind a outcomes those patients had, and obviously stratifying for PORT scores, and so

on, would be critical.

DR. BRATZLER: Yes. So remember the way this data set was selected. The charts are selected based on the patient population that had a discharge diagnosis, principal diagnosis of pneumonia, which means that after study, that was the reason they were admitted to the hospital.

Or they were admitted with respiratory failure, or sepsis, with a secondary diagnosis of pneumonia. That's how the case population was selected.

We chose to exclude patients who did not receive antibiotics in the first 36 hours, because we made the assumption that the physician probably didn't think they had pneumonia at the time of arrival. So that's why we excluded that population, just like we excluded the patient where there was no quote, working diagnosis. Those patients got excluded also.

DR. DANKNER: Okay; thanks.

DR. SPELLBERG: But to follow up on that, it does seem like that would be a very promising area to understand, quote, unquote, the effect of placebo because it sounds like those were patients that were probably misdiagnosed at first, and not treated as they should have been.

And maybe if we understood better how those patients did, we'd have a better grasp on the natural history of pneumonia, to some degree, in the modern era.

DR. COX: Dr. Mandell.

DR. MANDELL: Yes. I just have sort of a comment, question, but the point about mortality as an endpoint has come up a number of times, and my understanding of attributable mortality is that basically, if you did not have the pneumonia you'd still be alive; right? So there's more and more data coming out, that show that if you're, especially 65 or older, and you're admitted with CAP, your chances of dying of an infarct

and stroke and things like that are much higher in that they extend out, in some cases, to a year.

So you could argue, and to me as a physician, that is part of attributable mortality, because had you not had the pneumonia, you would not have had that MI or stroke.

DR. BRATZLER: I agree. I think there's similar data for acute influenza and other conditions, where--influenza's a great example, where most patients probably don't die of influenza-related respiratory failure.

They die of acute myocardial infarction, stroke, or something else, which may be closely tied to the influenza infection.

DR. WUNDERINK: Just one caution about using patients who didn't get antibiotics as a control group. They're a very different group of patients than the ones that we would put into a clinical trial, and

so their mortality, I would guess, is actually going to be lower than what you would look at for true pneumonia, where it manifests right from the very beginning.

As Dale knows, you know, the patients who get delayed antibiotics are actually different patients than the ones who get antibiotics from the very beginning.

So I'd be very cautious to say that's a good place to look for our, you know, baseline mortality of untreated pneumonia. Those are different patients. There's a survival effect there. They live three days to finally get their antibiotic. So, you know, I think that that would be--I'd be very cautious to use that group.

DR. COX: And then I'll just ask Dr. Fleming to make a last comment here, and then we'll move on to our next speaker.

DR. FLEMING: Great. Well, Dale, it's a very interesting database with a lot of insights that are emerging from this. I would

say this database, though, like others that are similar to it, provide very significant insights.

Those insights are in three areas. First, what is the overall rate of outcome that occurs? So as we look at a wide selection of 20,000 people, what would we expect to occur, and in particular, you're looking at from a mortality perspective during an inpatient period or out through 30 days.

And the second is what are the prognostic factors? What are those characteristics of people that put them at higher or lower risk for that mortality endpoint?

And the third is descriptions of how patients are managed, and then specifically here, you're able to characterize the different antibiotic selections.

Those issues are ones that are invaluable as you're planning a clinical trial and you're selecting patient populations and

you're planning sample sizes, to try to determine what is in fact the likely event rate.

Where, however, these databases are much more challenged is when you try to get a causality, when you're trying to say is this choice of antibiotic better than that choice of antibiotic.

It's descriptive evidence, it might generate a hypothesis, but when you're seeing, as is the case, as you would expect to be the case here, relative risk that are in the range of .5 to 1.5, selection factors can largely be accounting for those types of differences.

I would argue that while there may be uncertainties, in many cases, as to which is the right choice, caregivers don't make these choices in a pseudo random way, and essentially you have to be assuming that other than the adjustment for the covariates that we have--and in fact I always say the covariates that we have, that describe the difference

between one patient and another, is the "tip of the iceberg" as to what really makes those patients different.

And so the vast majority of what is different about them, or different about what their caregiver would decide, can't be adjusted for using even the most sophisticated statistical analyses.

So what I would say is these analyses aren't telling me what's the magnitude of difference that I could expect to see, or what is in fact the true magnitude of difference in the choice of these interventions.

So there are limitations to what we can learn here. What we can learn here, though, that is invaluable, is what is the expected event rate, how are people managed, what are the prognostic factors.

That's what the real strength of this type of database would be.

DR. TEMPLE: Can I understand one

thing. Experience tells you, in many outcome studies, that trying to decide what the exact cause of death is is very difficult, and so it's very common for us, in those studies, to look at total mortality or cardiovascular mortality, and breaking it down further is difficult.

In a non-inferiority study, of course, as usual, failure to do that introduces a bias toward the null. Very few of the deaths are because of the infection itself but are just because the person's old, and old means more than 75, by the way. I have to tell you that.

[Laughter]

DR. TEMPLE: Then you tend to declare equivalence. So it's more tempting than usual to try to actually figure out what the cause of death was and try to get the infectious deaths, and that's probably what the older studies showed, mostly. I mean, when you had 40 or 50 percent mortality, it

wasn't from heart attacks in most people. It was because the infection got them.

So anyway, as usual, the incentives are opposite.

DR. FLEMING: And Bob's, as usual, fully correct. It is more tempting than usual, in a non-inferiority trial, for the exact reasons that you indicated, to try to understand causality.

However, the reality is understanding causality overall is an extremely difficult thing to do, and in this specific area, for causality specific to pneumonia, and I fully agree with the discussion that was said earlier--even deaths that appear to be completely unrelated maybe aren't completely unrelated because of correlations of the conditions of patients.

So it is, in a setting like this, very difficult to be able to fine-tune with, because we don't really have the ability to get reliable causality data. However, with

that in mind, it makes your point exactly correct, and that is non-inferiority analyses are challenged when it may be the case, that in an earlier population that we would be using to assess the effect of, let's say, penicillin, or whatever we're using, we had a population that was more specifically cause-specific pneumonia, or cause-specific CAP.

DR. GILBERT: So I'm getting the cart and the horse feeling here. So Lionel's point's very valid, but if you have heart failure, you're more predisposed to the pneumonia. So, you know, which way to analyze this becomes very difficult to figure out, it seems to me. It can go both ways.

Well, now the question is if we get into subset analysis, if you will, if we go from empiricism to being able to identify the etiology of the pneumonia, what are the implications and the title of the next presentation is, Can we improve the detection of streptococcus pneumoniae? Keith Klugman

certainly is an expert in this area, professor of Global Health at the Rollins School of Public Health at Emory.

Keith, can you help us out here.

DR. KLUGMAN: I want to thank Dave and the chairs for inviting me. I hope to be able to help out because we have some real problems here, and I think that we do have tools which can certainly help in the setting of clinical trials. They may not yet be in clinical practice but in trials, I think we do have some answers here, and I want to sort of, before showing any slides, to think about the basic empiric idea that we have, that we think we can treat community-acquired pneumonia of various types of severity.

In fact when we're doing that, because we're using drugs that were designed specifically to kill bacteria, what we're actually doing is treating presumed bacterial pneumonia, and I'm really asking for some kind of a sea change down the line, that drugs are

going to be given and the labels are not going to say this is a drug for community-acquired pneumonia. This is a drug for presumed bacterial community-acquired pneumonia.

And why that's important is the sorts of things I'm going to propose here can be very frightening to industry because they're afraid that if they do a trial in which everybody gets a Binax positive test when they come into the trial, they're going to get a label, this is a drug just for pneumococcal pneumonia or proven pneumococcal pneumonia, and that's not the big market.

But down the line, I think we need to move to register drugs for bacterial infections or at least for presumed bacterial infections.

So with that rather long intro, I want to give my disclosures, and then I also want to show this one slide because I think it's really important.

I am, after all, at a school of

public health, and we are talking about the single largest cause of mortality, infectious cause of mortality globally. So this is an extremely important subject. It actually dwarfs mortality from all the major infectious disease areas that get all the money, from AIDS, TB and malaria, acute respiratory infections, euphemistically called--you don't die of acute respiratory infections. You die of pneumonia.

So pneumonia's the number one infectious cause of death in both children and adults.

Now many people have alluded to our problems. This is just another restatement of it. But I want to point out that we believe that blood culture identifies less than 10 percent of presumed pneumococcal pneumonias. So blood culture is the gold standard but it is so insensitive that it really is only a tiny fraction, not of community-acquired pneumonia but of community-acquired

pneumococcal pneumonia.

So there is, the slide essentially points out that all other attempts to define a pneumococcal etiology either involve kind of heroic measures which are not routinely used in clinical practice, but I want to look at some of them that have been picked up and perhaps discarded. I'll re-look at serological tests. I'm going to look at urine antigen and I'm going to look at issues of PCR.

Before I want to do that, we do have another tool. Unfortunately, you can't use it in pneumonia trials. But it has given us insight into some of the criteria we use at the moment for presumed bacterial pneumonia. So what do we do at the moment? We use an x-ray. That's the sine qua non for getting into a clinical trial of pneumonia in adults, because you have to have a positive x-ray.

And then we did get a little bit about CRP and procalcitonin.

Now to begin to tease out the pneumococcal fraction of pneumonia, we have this tool which is called a vaccine probe. So what do I mean by this? We now have a vaccine.

Unfortunately, all these initial data are in kits, because the 23 valent vaccine in adults does not reduce non-bacteremic pneumonia. But the vaccine that is designed for young kids reduces pneumonia in children, and therefore if you have a randomized trial in which half the kids get vaccine and half don't, you can begin to tease up, well, what kind of pneumonia this vaccine preventing? And as we assume, which is a reasonable assumption, that the vaccine only prevents pneumococcal pneumonia, it tells us something about what is pneumococcal pneumonia.

So an illustration of this is a trial that was conducted here, well, in California and there are two groups. This is

the group that got the control group and this is the group that got the vaccine.

And all clinical pneumonia, there's almost no difference in these two groups. So the definition of clinical pneumonia in children was not very good at picking up pneumococcal pneumonia, because kids who got a vaccine that prevented pneumococcal pneumonia had almost no difference in the two groups.

Here is the utility of an x-ray, however. If a radiograph was obtained--this is just that they asked for an x-ray. So they were suspicious enough that it may be bacterial. Suddenly, there is actually a 10 percent difference in these groups.

And then if it was read as consolidation--and this is a huge trial in 40,000 kids and there were more than 300 radiologists reading these--there was a 20 percent difference in the two groups.

And then finally--and you'll see this is a later reference--they actually,

because of the electronic database, they were able to go back and take all of these x-rays, and get them read by a panel and there's actually a definition for consolidation on x-ray all read by the same radiologists.

Now there's a 30 percent difference between the two groups. So what this is saying is that if you get the same people to read all the x-rays, you can come up with around 30 percent of pneumonia that is x-ray confirmed, seems to be due to the pneumococcus, and that's beginning to get into the ball park of where we think the pneumococcus is playing a role in pediatric pneumonia.

Now this has been used in a number of settings, and I'm only showing this slide because in Gambia, which is a rural part of Africa, a very similar vaccine now reduces radiologically-confirmed pneumonia by 37 percent, so an even higher percentage, and just getting to the last discussion, it also

reduces all-cause mortality by 16 percent.

And you try to work out what patients have died of in an American hospital. That's a challenge. You try in rural Africa; it's impossible. So they specifically chose all-cause mortality because of the confounders involved in trying to get attributable pneumonia, which is not appropriate.

But what I want to get to now is to say, okay, that suggests that x-ray does enrich a population for pneumococcal pneumonia. This is now a South African trial which we did, same vaccine, and again we're looking at two groups--in fact this is the group that got the vaccine, this is the group who didn't. A 20 percent reduction in x-ray confirmed pneumonia.

And you can work out the fraction. This is the total reduction in pneumococcal disease, then, based on that 20 percent reduction, of 100 episodes per 100,000 immunized kids.

If you go up here, there is 2.5 times that burden of disease prevented by the vaccine, if you go to clinical pneumonia, beyond x-ray confirmation. So let me get into that a little bit further.

This is alluding to CRP and procalcitonin. So there's the WHO-confirmed pneumonia. This is the percentage efficacy of the vaccine in intent-to-treat analysis just of x-ray-confirmed pneumonia. When you add in a very high CRP, you can begin to up that fraction. Very high procalcitonin, you can up that, and the combination of the two, you begin to get even higher.

What you're doing here, however, is you're getting more and more specific, and reducing sensitivity. Another way of looking at this.

If you take kids who don't have an x-ray-confirmed pneumonia, so these are kids with no consolidation on x-ray, in both HIV and uninfected, and these are infected and

uninfected kids, vaccine efficacy is nothing. So if they don't have an x-ray, you say, well, there's no pneumococcal disease out there because they didn't have an x-ray and there's no protection whatsoever.

But they're a large group of kids, and if you do a C-reactive protein and it's greater than--that's 40 mgs. per liter, 4 mgs. percent. If the CRP's greater than 40, suddenly, you preserve all your protective efficacy of the vaccine, it's up to 32 percent.

So there's a large fraction of cases out there that don't have x-ray-confirmed pneumonia, but if they have any infiltrate on x-ray and array CRP, they have--they're protected. So the vaccine attributable reduction in disease is pretty much 350 versus around 134, just for x-ray alone.

So what am I getting at here? What I'm saying is that yes, the x-ray does define

a group of individuals who have pneumococcal pneumonia but there's a large fraction of vaccine-preventable pneumonia, at least in children, which is not x-ray confirmed.

So it's not going to help us inordinately, but CRP and procalcitonin do help to identify these. So it's basically a plea, that perhaps CRP and procalcitonin can be added to algorithms.

Now let's get to the specifics of pneumococcal diagnostics. The first hope was PCR, and I'm afraid straight PCR was a disappointment.

There were any number of studies looking at PCR in blood, and what the summary of this is is that essentially it was less sensitive even than culture.

And there was the one problem. The second problem was that in kids, it was totally useless. If you look now at non--this is a non-quantitative PCR. Little promise for the diagnosed pneumococcal pneumonia in

children or adults, and here's a study looking at PCR in the blood of control children and adults, these are not with pneumonia, and this is the percentage that were giving positive reaction.

So this is by age. So in kids it was totally useless, and the idea is that, in fact, carriage in kids will give you a positive PCR in the blood.

Now there is a caveat to all of this, and this was using a pneumolysin-based PCR. A pneumolysin, although it's a hallmark and it's important in respect to the pneumococcus is also found in some commensal streptococci, and some Alpha streps, and so on. So there could be a non-specificity issue here.

So in looking for a target, there are a number of pneumococcal targets that have been looked at. This is a big CDC study and it turns out, now, that lytA--this is the amidase protein, this is the protein that the

pneumococcus uses to commit suicide. This seems to be the most specific, and most of the newer data are using lytA as a target for PCR rather than the pneumolysin, and PsaA I'll talk about as well.

So this is probably the most important slide about the molecular diagnostics that I want to show, and this is the promise. Now it's not ideal for clinical practice yet but certainly is the "wave of the future."

So real-time PCR, which gives you a quantitative PCR, is more sensitive in culture, especially in patients receiving antibiotics and there have been a couple of studies so far, and essentially, the premise here is that patients, we have to separate colonization from infection, and it seems that the quantitative nature of real-time PCR, you can set a cut-off, and above that cut-off there is some indication that you may be looking at pneumonia.

So some of these types of studies is just an illustration.

Here is a cutoff of 10 to the 4 organisms per mL, and you're looking here at nasopharyngeal aspirates, and so these are culture positive, and PCR positive, and that's the range in patients with pneumonia.

There's a lot of control patients here, and there's a big block, and these are all the negatives. So there's quite a difference between this group with pneumonia and this big group down here without. There are a couple that are culture negative and are still positive with the lyt PCR. You're cut off, it's not a whole lot, but there are some of them, and then culture negative, lyt PCR negative. You know, these theoretically are those that are not pneumococcal pneumonia.

Is there a relationship between these quantitative real-time PCRs and outcome? So far the best data only come from meningitis. This study did try to look at

pneumonia but there are very few pneumonias in it, but they showed very clearly that in meningitis, high levels of real-time PCR in the CSF were associated with increased mortality.

The other place the real-time PCR is coming into its own is in empyema. It's a small fraction of the diagnostic group but it's important, and there are a number of studies now that culture negative empyema. This is now trying to culture the pus from the empyema. Most of them have had antibiotics already before anybody stuck in a needle. Sent for real-time PCR. 75 percent positive. Okay. And in fact many of, half of the positives in fact were serotype one, which is another whole discussion.

The PCR was using the pneumolysin gene, but you don't expect to get streptococci floating around in pus fluid.

Here's another complicated picture of a whole series of patients, but the message

is that in many of them in whom pleural fluid culture was negative, a whole group over here, the pleural fluid PCR was positive, and, in fact, pleural fluid probably, in a very short period of time, will routinely be sent for PCR.

Just for this particular audience, there is a future which may have nothing to do with finding the pathogen, or it may have to do with differential gene expression.

So these next two slides are just conjecture for the future.

The differential gene expression thought goes as follows. Pneumococcus that is in the nasopharynx is expressing a whole repertoire of genes which are related to colonization.

As soon as it gets into the blood or into the lungs, it produces a completely different repertoire of genes, and we're able now to measure these things with microarrays.

There are issues of sensitivity.

But eventually, we may be in a situation to say, okay, this is a pneumococcus, this is its repertoire of genes it's expressing, and therefore this is an invasive pneumococcus as opposed to a colonizing pneumococcus. So that's just or the future.

And then the second for the future is that we may not have to identify the pneumococcus at all. With protein genomics these days, there is a host response to pneumococcal disease, which is different to the host response to malaria, which is different to the host response to TB.

And simply if you have thousands of host proteins, which may be, in fact, all measurable already now, you could, with good mathematical programs, work out a host response to a pathogen without finding the pathogen at all, and this may be the future of diagnostics where, in fact, host protein arrays are specific enough to give you a diagnosis.

Okay. Enough for the future. Let's go back to now, two more things that can be used in clinical trials, and I think both of which outweigh what I've said up to now.

The first is a re-look at serology. Now serology is not useful in clinical practice. It's not particularly useful for you, in retrospect, to know that you had a pneumococcal case, if you have to wait for two weeks or three weeks to get a change in antibody concentration.

But in a clinical trial, I think it's an underused modality. Clinical trials, we routinely bring all of our patients back for follow-up visits, and there is no reason why serology can't be taken up front, and then a follow-up.

So the most promising in terms of the pneumococcus is PsaA antibodies. This is a study from Kenya. Sensitivity and specificity of a 1.3 fold increase in antibody. Unfortunately, it is not useful in

terms of a single diagnostic. You can't use this in a "one off" acute serum, but if a patient has pneumococcal pneumonia, there is this study and there are a couple of others on the next slide, all of which suggests that there is around a 1.3 to twofold rise in antibodies, and sometimes greater.

You can't see these numbers but there is around a 20-fold, 25-fold increase in the antibody titers to this PsaA.

So it's just a thought, but if you are bringing patients back, and you want to end up with a specific population in pneumococcal disease, serology is not entirely out of the window.

This is just a little illustration, again, of exactly the same serology. This is an outbreak of a pneumococcus Type 4 in a nursing home. So there were 18 pneumonia cases from one long-term care facility, hospitalized over a two week period. That's a classic example of an outbreak.

How good was blood culture? Well, three of 18 pneumonia cases had a blood culture. They were all the same, top four pneumococcus. So I mean, it is at least suggestive that all of these 18 cases may have been due to that pneumococcus. If that's the case, six of them had additional twofold rise in this PsaA antibody. One had a pneumococcus in culture from sputum and one patient had a latex agglutination.

Okay. So the "gorilla in the room" here is the Binax test. Unfortunately, again, for John Bradley, not good in kids. But in adults, this is emitter analysis of the sensitivity and specificity, sensitivity around .74, so pretty good sensitivity to diagnose pneumococcal pneumonia.

Of course what's the gold standard? This is using a basket of any number of existing tests, sputum and many others, to try and define the group that you base the, compare the sensitivity to. Specificity very

high, though, .94. So if it is positive, it seems to be a useful kind of a test.

So what is this test? For those who don't know, this is just a test of urine, a very quick test which is looking for C-polysaccharide in urine.

Now it's not a perfect test. I'm going to go through some of the issues around the sensitivity. You can increase the sensitivity by concentrating the urine but that then defeats the utility of the test, because what we're really looking for is something that can be done as a dipstick immediately before enrollment in a trial or before presumptive antibotics.

So I'll deal with some of the issues about that. And then there are some other issues about how long does it stay positive, and so on. So I'll go through some of these things.

So specificity, as I said the specificity is very high, the limitations are

we don't know what the gold standard is.

So conventional microbiologic methods are such, that we really don't know the difference between false positives and true positives. If this thing is better than the existing, is it a false positive or is it really a true positive. Some of the issues around that are shown on this slide. I'm not going to go through them all, in detail, for time.

The issue of antibiotics. This is quite a nice study, looking at patients who didn't have prior antibiotics. In the vast majority of those then is a positive sputum culture and a positive Binax.

There is a group, and these are the ones we don't know really what they mean, these are the positive Binax with a negative sputum culture, but it may be that if the test really is specific, that those are true positives.

Once you have antibiotics, of

course the sputum is gone, and then you get another group of positives over there. So that's the issue of previous antibiotics. I'll come back a little bit more to it. A worrying thing is this issue of persistence, and there hasn't been, I think, enough study on this.

A couple of studies now, these are all different references, looking at greater than six days, day seven, four weeks, even up to six weeks. And finding positives. So in many clinical trials, you can't get enrolled in the trial if you've had a recent previous episode that may deal with this, but there may be some issues about how long this test stays positive.

If you have a concentrated urine, then you have more problems perhaps with that assay.

Sensitivity issues. Well, this is not a very, very sensitive assay. As I said, you can improve the sensitivity, I'll show you

the data, if you concentrate. Even the reading of it is a little difficult. It's just a line that you're picking up and most of the tests now would say that any positive signal is a positive. So the test could be made more sensitive.

Is there a higher positivity for severe disease? There are a couple of studies that seem to go in that direction. This is one of those. Non-severe CAP and then severe CAP; These two patients died.

And a little bit of an indication that when you titer out the urine, there is more positivity and more severe disease. I guess the idea is more burden of bacteria, therefore giving a better outcome. The issue in relation to antibiotics with this assay is confusing, to say the least. There are data both ways.

So higher positivity after antibiotics and lower after antibiotics, and I think what we're dealing with here is a

bimodal curve.

As soon as the organisms are killed by the antibiotic, there is antigen released and there's probably more antigen in the urine. Therefore, if you get it just at the right time, after the exposure to the antibiotic, this test may be more positive.

Once you've killed the bacteria and the bacterial load drops, then, over time, the thing drops. I think that's what's confounding this but we really don't have enough good data on exactly what the role of antibiotics is in making this positive or not.

This is the concentrated urine issue. In general, a study with a lot of different groups of patients. Sensitivity improved from 27 to 38 percent, if you concentrate the urine. This is definite pneumococcal disease, blood pleural culture form 75 percent up to a 100 percent probable, just a positive sputum from 44 to 69. So you see there are some sensitivity issues with

this which can be improved by concentrating the urine.

In a context of a clinical trial perhaps that could be done. It defeats, unfortunately a lot of the utility of the assay, if you have to get around concentrating at the time, will take a lot more, and concentrating is not inexpensive either.

This is a further inference from this, I'm not going to go into it in great detail, but they were in this study trying to get at the idea that if you're using beta-lactam monotherapy, perhaps where mycoplasma is identified by PCR, you can then look at the Binax and see if it was positive or negative, and so if you have a positive Binax, you would expect that in cases of a positive Binax, here we have patients with a positive Binax, here, that's a negative Binax, and where there is a mycoplasma identified, the idea would be that the Binax was negative and the penicillin monotherapy would do badly and in fact only

six of fifteen responded.

But that's getting beyond our arena today to use the Binax as an indication of outcomes.

So in conclusion, where are we at? The conjugate vaccine probe study suggests that vaccine-preventable pneumococcal pneumonia, at least in children, extends beyond classical lobar consolidation. There's a burden of disease preventable beyond x-ray confirmed pneumonia, and that CRP adds value if there are other changes on x-ray, and so does procalcitonin.

So while PCR on blood has been disappointing, real-time PCR on sputum, or even on nasopharyngeal aspirate may be promising.

So we've dismissed for years, and we've had many arguments about what is the value of sputum. Well, it's heresy to put up that a nasopharyngeal aspirate might be useful.

But the idea here is that if carriage represents in adults a very low, in general, a very low level of colonization, and that pneumonia is associated with a much higher level of colonization, that maybe this quantitative PCR, even on nasopharyngeal aspirate may be promising, and some big study's going on at the moment.

For the future, proteomic studies. For the moment, then, what can we propose for clinical trials? Binax, certainly, I think, is likely to be useful in clinical studies in adults.

I don't see any reason why you can't do a urine Binax on everybody in a clinical trial and enroll those that are positive.

Serology using PsaA with paired sera adults may be a useful adjunct to diagnosis in pneumonia studies when you're bringing the patients back anyway. Thanks for your attention.

DR. GILBERT: Thank you, Keith, very much. We have time for comments and questions.

Yes, John.

DR. BRADLEY: A spectacular review, Keith. I realize how little I know every time you get up and present your summary of the field.

In trying to take the information that you presented and directly apply it to my enrolling patients, and when I see them on the ward, or my adult colleagues see their adult patients on the wards, rtPCR certainly holds promise, but to have a lab with an rtPCR machine that is available 24 hours a day, because you want to enroll them as soon as they hit the wards, to negate this previous antibiotic effect, is hugely expensive and ostensibly all of the support will come from industry, which is putting money into allowing the investigators to collect the data.

There were some presentations

yesterday on gene chip arrays, and an approval of a device for viral infections, and I know there's some work being done for bacterial infections, and to have a gene chip that you could run on a patient is a technically easier test to do, takes less time and expense, and may be more practical.

And I was wondering if you could comment on where that part of the field goes.

And an interesting observation. Children have all of these false-positive rapid diagnostic tests for pneumococcus, and somehow, at some level, I believe that actually these kids have early true infections at the mucosal level, particularly infections that are both viral and bacterial at the same time, and what we see is actually an early infection that the host, these young excellent immunologically-competent hosts are capable of addressing without need of antibiotics, so they get better without progressing to the classic pneumonia but they're truly infected.

So they're not really false-positives. So it's how you define false-positive. So could you address the gene chip?

DR. KLUGMAN: Two good points. I'll address the second one first. I agree with you. There was one quite nice study looking at, prospectively at kids, and when kids are newly colonized with a pneumococcus, they often have some signs and symptoms, and almost like just a mild respiratory illness, and that may be what you're talking about.

In terms of the usefulness of the Luminex-based platforms for microarray, unfortunately for the pneumococcus we're going to come up to the same problem. The microarrays at the moment are qualitative rather than quantitative.

So if they pick up pneumo, I fear that in kids they're going to be picking up this carriage signal as well.

It may be possible to make them more quantitative over time, and that's

something that we're actively looking at getting involved with. But in adults, they may well be more useful. But again, it's this quantitation issue. It seems to be in, for, certainly for the bacterial pathogens, where we're accepting there is a carriage state, that you're going to need a quantitative element to them, if they're going to be more useful.

DR. GILBERT: Dan.

DR. MUSHER: Keith, I think I didn't understand. You're proposing that in studies of pneumonia, that we try to separate out pneumococcal pneumonia cases at the start, and treat them under a protocol. There might also be studies of pneumonia without including pneumococcal patients, or all-comers because you don't try to distinguish pneumococcal pneumonia?

Would you help me with that?

DR. KLUGMAN: Well, what I'm trying to get at is a group of patients who have

presumed bacterial pneumonia, and you could come up with some kind of idea of how to do that, but one way may be simply to do this test, and then you're recruiting a group of patients who have, with relatively good presumption, pneumococcal pneumonia, and we all agree that the pneumococcus is the number one pathogen that you're trying to treat in all of these.

Now there are going to be some patients who won't have a positive test, who you're still going to include in a definition of pneumonia, and that gets tricky, but you could say we're going to take everybody with us, kind of an x-ray or whatever, or this kind of a severity score. But at least you're going to greatly enrich, I believe, the pneumococcal population in your trials, and hopefully you can then do analyses which are restricted to the pneumococcal population, which might have more rational kind of outcomes.

I mean, it really worries me, that if these trials are moving towards these PROs as outcomes, that is fine as long as you're assured that the group that you have enrolled have true bacterial disease. If they don't, then it becomes meaningless, because I can conceive of a study in which I gave a mood-enhancing drug to a group of patients who had pneumonia, which is going to resolve spontaneously, and they would all feel better, and suddenly this would get a license for pneumonia. I mean, that's total nonsense.

So essentially we have to define a group of patients who have bacterial disease, because the drugs are designed to kill bacteria, and this may at least be one way of enhancing that microbiologically valuable group, where you can then look at any kind of outcome measure that's useful down the line.

DR. MUSHER: I would like to comment, just from the point of view of those of us who are treating patients with disease,

I think that now that we're able to--I think we're going to be able to move away from this very restrictive four-hour treatment in the emergency room, the IDSA/ATS guidelines no longer list that four hour, and my understanding is that the JCAHO is going to be asked to remove--David, is that correct?

DR. GILBERT: Well, we should ask Dale. There he is. My understanding is that there are several provisos now included, and there's diagnostic uncertainty, is one thing that is often quoted.

Dale, do you want--

DR. BRATZLER: Yes.

DR. MUSHER: And then I'll come back to my--

DR. BRATZLER: So the measure actually is six hours now, not four hours. It was officially changed to six hours.

DR. MUSHER: That's what I thought.

DR. BRATZLER: Last year. And then the other thing that happened is patients who

don't have a diagnosis, or if the ED physician documents that the diagnosis was not clear at the time of arrival, those cases are excluded also.

DR. MUSHER: So six hours gives you the time--I'm now speaking not as someone who's testing drug but taking care of patients. Six hours gives you the time to get--for sure, to get a urine antigen test.

DR. KLUGMAN: See, what worries me is that when we saw yesterday the data of who's actually in these clinical trials, there are vast numbers of people who are in PSI one and two, and I think you could say that if you're in PSI one, and two and you have negative procalcitonin and you have a negative urine antigen, you're not in the trial.

That could over time change practice, so that eventually, if these things are available, we'll begin to start using antibiotics only when we've got better indicators of a bacterial etiology.

DR. MUSHER: Speaking as a practitioner, if you are in class one or class two, and you don't have the specific evidence for a pneumococcal infection, then you absolutely should be treated with a macrolide, or tetracycline, and the whole problem is--and I commented yesterday on the Swedish recommendation--if you think that that group is going to have a pneumococcus in it, I'd rather, if I thought the patient had pneumococcus, even though it was mild right now, I'd rather give the penicillin, still coughing, in a few days I can reconsider the mycoplasma, then do it the other way around.

And I think again with regard to who we see coming to a hospital emergency department, I think if we have a little bit more time, we should--we've talked about it. It's only been inertia, Keith, that we haven't done it at our hospital. We should routinely be doing that urine detection test of a pneumococcal antigen. It's very specific and

it's quite sensitive, and it's sort of crazy not to be doing it.

And a little extra time is helpful. And the same thing with analysis of sputum for Gram stain. I wanted one further comment, because it hasn't been pointed out. One of the reasons for all the problem with the delay in therapy, and we infectious disease doctors were part of it.

You get the interns, they say, well, I haven't yet got a sputum on your patient, so we'll just wait, try a little bit later on. And it would get to be four hours and six hours, and eight and twelve and sixteen, and at a certain point you're sorry you didn't go ahead and start the treatment.

DR. GILBERT: You want to turn your mike off, Dan, please.

DR. MUSHER: Sorry.

DR. GILBERT: I'm going to ask Tom--

DR. MUSHER: Was that your polite

way to ask me to shut up? Or just turn off the microphone?

[Laughter]

DR. GILBERT: You're doing great. I'd ask Tom to comment, sort of on an initial basis, and he may have more comments later, after more presentations. But obviously there's a moving target here with these modern diagnostic techniques, and Keith has given us a wonderful review of the current status. So if I'm trying to design a clinical trial now, and I know at the outset, that I'm going to have a subset analysis in my clinical trial, I may not be able, at the outset, to say that I've got pneumococcal etiology, but using these techniques, I know at the end. I'm going to have a substantive subset that's going to be either pneumococcus yes, or pneumococcus no.

So how do we integrate that, prospectively, into clinical trial design, so that we end up with up with valid power for

interpretation of results?

DR. FLEMING: So just to give a preliminary answer to that, because I think there's more richness to all this that I'd like to see play out today as the discussions go on.

If you have, in advance, as you design the trial, a very clear-cut way of defining who it is that is your ideal subgroup or your targeted population, that group that would have what you call pneumococcal etiology. Hopefully, in designing the trial, we're going to be able to be sufficiently selective, that that group will represent a substantial fraction of the entire trial.

If it does, then certainly you can build in a prespecified subgroup to say I'm going to do my principal analysis, or I'll do one of the analyses in that group, and obviously to be powered then you would have to have adequate numbers of patients in that group.

So if we were looking at whatever endpoint, if it was a mortality endpoint, ruling out a given non-inferiority margin, then you would ideally want that subgroup to be a substantial fraction of the entire group.

If it was two-thirds of the entire group in the end, then you would need three halves of the sample size. Of course the other part of the complication here is you can't ignore the other group when you're looking at overall safety issues and benefit-to-risk issues.

Where it becomes a lot more complicated, or even more complicated, is when you can't tell me, in advance, what is the exact characterization of that group of interest, and then we start exploring the data to find those groups where it looks like the signal is the best, and then we define that to be the target group, and we can talk more this afternoon about why that leads to great risk and misinterpretation.

DR. GILBERT: My other point is--a quick follow-up. We're better off to do it prospectively rather than have these retrospective analyses that always raise everybody's hackles.

Bob.

DR. TEMPLE: It's only retrospective if you define the group after seeing the data. If you specify--I mean, if there's a test that takes three weeks before you know that it was really pneumococcal, it's perfectly okay to do the analysis in that group. I mean, the overwhelming tradition in antibiotics is to start treatment and then see if they have a sensitive organism. I mean, they've all been done this way, for years, and I don't see any real impediment to that, as long as you identify it prospectively, because it's a baseline characteristic.

You of course can't--I mean, it's not a stratum you can randomize to because you don't have it identified. But if the trial's

big enough, that's usually not much of a worry. I think one should assume that. The main question, though, you have to decide, is how what you're doing affects this population compared to the population you have results data in.

You know, if it was an unselected population in the past and now you're using a selected population, are you confident the effect of the control is at least as good as it was in the past? If it's better, that's okay, actually. That probably increases the strength of your study and it probably would do that if you got people who definitely had a susceptible organism.

DR. FLEMING: So I think we're saying the same thing. There are three or four key issues here. The first is it is far different when this is a well-defined, pre-specified algorithm as opposed to something that you define as you explore the data.

The second is if this group is not

a substantial fraction of the entire population, the efficiency of the design is less because you are leaving out all those people that aren't meeting your definition. The third is that there are still safety issues in benefit to risk, so you can't ignore the people that turned out not to have the pneumococcal etiology.

And then as Bob says, in the end, if you're going to do a non-inferiority analysis, what we have is historical evidence or historical trials, and if they didn't use the same population, then there are issues about the constancy assumption and how that impacts the non-inferiority margin.

DR. GILBERT: Dr. Rex.

DR. REX: Well summarized. I want to pick up right where Tom stopped talking, because I think there's a theme here that we might be able to take advantage of. We spent yesterday recognizing that we don't have placebo control data, that it's going to be

very hard to get placebo control data except perhaps in a very carefully selected subset of very well young adults who we believe could tolerate not being treated for a period of time, and they don't generalize to the 65-year-old that started today.

So I don't have placebo control data, I don't know, I don't know, and I've got constancy issues. But maybe the question of assay sensitivity that sits at the heart of whether or not you can--part of the heart of whether you can believe a non-inferiority trial is fixed by what Dr. Klugman was talking about.

Because all this diagnostic stuff--I was for a while thinking the diagnostics don't help me, because all it does is make it more certain that the patient needs antibiotics. I know it's the pneumococcus. I really think I ought to treat him.

Maybe that's something that actually helps us here, because if I now say

I'm willing to only study pretty sick people and they've all got to have a procalcitonin of five and a CRP of 120, and maybe they don't all grow the pneumococcus, but they looked like they probably had a bacterial cause, even though I don't, I can't justify a lot of the stuff that I'd really like to be able to justify--I really would but I can't--maybe that helps us buttress our concerns, our angst about assay sensitivity.

So I just want to point out the theme, the way that you can use diagnostics to get at one of our key points of concern that I think we're going to debate later.

DR. GILBERT: I think that was a comment and not a question specifically.

DR. TEMPLE: Let's say you have some past data that make you reasonably sure that treating bad pneumococcal pneumonia was good for you. There was a reduction in mortality of something, I don't know, 20 percent, whatever. We're going to hear data

later. You worry now if it's still relevant to the present population.

I think you only gain if you get greater assurance that the people in the trials now really have pneumococcal pneumonia. It makes you think that past estimate has some validity. Whereas if you're not sure how people are diagnosing, or maybe they're treating at the drop of a hat now, and they didn't used to before, that would undermine your constancy.

So I think anything that makes you more sure that they have--you know, they have to have the appropriate degree of illness and all that. But anything that makes you more sure it's pneumococcal should enhance your feeling, your assurance about constancy, I would say.

DR. KLUGMAN: I want to support that, and then also point out that the demographics of pneumococcus disease have changed dramatically. In effect, the host

factors are 99 percent now your risk for mortality in pneumococcal disease, and we have a lot more people at risk, and the whole spectrum of pneumococcal disease has changed.

So much more susceptible individuals now with underlying illness are getting pneumococcal disease. So for this constancy argument, I would argue that perhaps, if anything, in the absence of any antibiotic, our populations that get pneumococcal disease today, one could argue would be at greater risk of mortality than they were before.

DR. GILBERT: Dale, I want to stay on time but--

DR. FLEMING: Just before we leave this point. But it's more than a greater risk of mortality. It's specifically attributable risk to pneumococcal, and so I would agree with the comments that have been made. Anything that would enhance the attributable risk ought to be something that should be a

reassurance relative to the constancy assumption, realizing that there are many other factors too that influence the constancy assumption.

DR. GILBERT: Okay. Dr. Powers, if you're really quick.

DR. POWERS: Three quick points.

DR. GILBERT: Three doesn't sound good to me.

DR. POWERS: One. There was actually more certainty of diagnosis in the past. The majority of people in the older studies had positive blood cultures for pneumococcus. So Bob's right--this will actually assure constancy.

Two. You can only do these subgroup analyses on baseline data that are captured at baseline. So doing something like a person having a persistent blood culture on therapy, you can't analyze those subgroups cause it's on therapy.

And then thirdly, the idea of test-

-this number of isolates thing, where we got ten of this and ten of that, those are exploratory analyses and really don't allow you to make any kind of confirmatory conclusions about effect of drug per organism.

But that's what we've been doing for a long time and it's getting to this, you know, counting up how many of this and that, and that really doesn't help you make confirmatory conclusions in the end.

DR. GILBERT: On that note, we'll start again promptly at 10:35. Fifteen quick minutes.

[A recess was taken from 10:20 a.m. to 10:45 a.m.]

DR. FLEMING: Let's reconvene, and we've asked John Powers to present, as we begin this session now, some key insights about primary and secondary and composite endpoints.

John.

DR. POWERS: Okay. Let me just

get started. In the interest of time, we'll speed through this again. I'd like to talk about today, some just general points about what does it mean to do any kind of measurement? So in this case, we're talking about measuring outcomes in a clinical trial. But what makes a good measurement? Then talk about some definitions that actually come from ICH guidance, about what are clinical endpoints and biomarker slash surrogate endpoints, and what is a primary and a secondary endpoint? And then finally talk about how do we analyze these endpoints?

How do we look at a single endpoint? How do we look at combinations of endpoints? What are the issues when we want to look at multiple endpoints with multiple testing? And multiple testing leads us to the issue of subgroup analyses and it dovetails quite nicely into the discussion that we just started to have.

So the first thing we want in an

endpoint is we want validity, and Dave Gilbert touched upon this yesterday. It means that the endpoint actually measures what it proposes to measure. Every measurement outcome has this. Clinician-reported outcomes. Patient-reported outcomes. Obviously mortality has validity all by itself. All-cause mortality. Cause-specific gets up did I really measure what I thought I measured?

So there's three things that go into validity and there's a great 17-page little booklet called Reliability and Validity Assessment by Carmines and Zeller that explains all of this in 17 pages. Worth a read.

Concept validity is does the measure capture all the relevant domains of what I'm intending to measure? For instance, if I want to measure kids' ability to do math, if I just ask them addition questions I haven't measured everything I need to know about their ability to do math.

By the way, I want to mention, when we all say we don't like PROs, anybody that's ever taken the SATs and the MCAT, that's a PRO. Right. All it is is attempting to measure an abstract concept in a standardized way. So we're trying to measure your intelligence or ability to get through medical school. We give you a standardized test, and that's all we're doing, is standardizing the measurements.

So essentially concept validity then talks about what we're going to measure. Construct validity is how well the instrument measures what it's intended to and how it fits together, how the pieces fit together. So that talks about how we measure it and when are we going to measure the outcome.

And then finally there's criterion validity. If we're coming up with a new measure, we want to compare it to some other things that we know essentially measure the same thing, maybe not in as reliable or

precise a way, but that's what we're trying to do when we develop an outcome measure, is make it more precise than what we had before.

Face validity isn't validity. It's not a scientific measure. It's eyeballing it and saying, oh, it looks good to me. But that's not a scientific measure. And the other issue is there is no such thing as validity. Validity only applies to the situation in which it was studied, which is the entire issue with non-inferiority. You can't take a measurement from my lab, where I used X reagent, take it to your lab, use a completely different reagent, and expect the experiment to come out the same.

So we want to make sure the validity applies to the way we studied it. Once we know what we're measuring, then we move on to reliability, which is that the measure is reproducible over time, and between and within observers. And I showed you some data yesterday that clinician judgment is

certainly not reproducible between and within observers.

But remember, reliability doesn't mean anything if we don't know what we're measuring in the first place, because we can get a precise measurement but we might just be measuring something that's more precisely wrong.

The reason why we want reliable measures is that means less variability and less variability means smaller sample size and ability to show a difference with fewer people.

The next thing we want is responsiveness, and that means that the measure is capable of detecting a change if a change exists. It doesn't mean we pick a measure that's going to change, even if it's not clinically meaningful.

The reason why you may not see a change between drug X and drug Y is because there isn't a difference between them, not

because the measurement scale was wrong, and I think that's something important to measure.

If people get better in eight hours and we have a scale that's responsive to change within eight hours, and you can't show a difference between your drug and placebo, it doesn't mean the scale is wrong. It means your drug's not having an effect.

And the last thing we want is acceptability. Responsiveness also brings up the issue of how much of a change is actually meaningful for people. If I could sell you an air conditioner that cools the room .00001 degrees Fahrenheit cooler than another air conditioner that costs $500 more, would you buy it? No. Because you can't feel that difference in temperature; it's not relevant to you.

And then finally there's acceptability. How can we get the information? And that applies upon missing data as well. What is a clinical endpoint? A

clinical endpoint is a direct measure of how a person feels, functions or survives. Feels is not warm and fuzzy. It's not scotch and soda, and therefore I feel better about the traffic on the Beltway. What we're talking about in a disease is the symptoms that are relevant to that disease. Now in depression, it does have to do with how you feel. But in pneumonia, we're talking about what are the symptoms that are referable to pneumonia.

I can't tell if a person feels short of breath. I have to ask the patient whether they feel short of breath and get that information from them. But that's not going to be something that we're just going to make them feel better by letting them snort a line of cocaine. It relates to actual measurements of the symptoms of disease.

A surrogate endpoint is defined in ICH-E9, is an indirect measure of effect and it actually says it should be used in a situation where direct measures of clinical

effects are not feasible or practical.

Well, we've talked a lot about, the last two days, about how people get better rather quickly in pneumonia and we can measure clinical effects directly.

Therefore, we have to ask the question of why would we need a biomarker or a surrogate variable in a setting like this. This is not HIV. This is not hepatitis where the actual clinical events may happen months to years down the line. We're talking everything happens in the short space of a couple of weeks.

When you want to look at this, the average duration of therapy in the early studies of community-acquired pneumonia was 107 hours. They got about four and a half days of therapy, and the average response time was two days, and Max Finland says, "And we gave them two more days just because we felt like we were going to do it."

So it also brings up the question

of why are we giving people 10 to 14 days of therapy when they were originally given four and a half in the first place.

And Rich has written an editorial on this recently, about why are we doing this to people. So when you actually look at this, these are not studies, these are individual descriptions and case series. But when you look at it, it's rather informative. This was the typical natural history.

What was the endpoint in these trials? I can't tell you because they looked at all sorts of stuff--pulse, temperature, blood cultures, acute symptoms.

By the way, blood cultures got negative after one dose of penicillin. So the idea that a single dose of therapy doesn't have an effect, which is also now bolstered by the analysis that the folks at Cubist did of pretherapy actually did have an effect on outcomes as well, at least in a post-hoc subgroup analysis.

But it's interesting, look at this first patient, that's day eight, nine and ten, when things are getting better. The other person on the right, well, that's day two, when things were actually starting to get better, and you can actually see--they give you a lot of information.

But how did they define "cure" here? I don't know, because they just give you all these pieces of information, which is very informative from a descriptive point of view.

But how would I take this and use this as an outcome measure in a future clinical trial? Don't know. So the other issue is: What's severe disease? Well, here's a guy that had two lobes involved you can tell, and it takes this person longer than the other two to get better.

So yes, pneumonia is a continuum of disease, but we can categorize it just like we categorize age and other continuous variables,

and that is that it appears that people with severe disease take longer to get better, and also may have a higher mortality.

It's interesting, when you look at this, that they said on the whole, these patients they evaluated in this case series represented severe disease by all the usual criteria.

So this is kind of those like "we knew it when we see it." But then they go on to say more than two-thirds were over 40 years old, the majority had two or more lobes involved, and appeared to be clinically ill, severely ill, with delirium, evidence of peripheral vascular collapse or congestive failure.

Notice they pointed out that the ancillary things like going into congestive heart failure were part of the disease process, not separate as Dr. Mandell pointed out as well.

More than half had positive blood

cultures. Somebody informed me that I misspoke and said every trial in the past had a high rate of blood cultures.

If I said that, I misspoke. But this one certainly did, and a number of the other ones have a higher rate of blood culture positivity than we see in current trials.

What's the primary endpoint? Well, ICH-E9 defines a primary endpoint as a variable capable of providing the most clinically relevant and convincing evidence.

It generally should be only one primary endpoint, and it states that it should be sufficient evidence that the primary variable can provide a valid and reliable measure of some clinically relevant and important treatment benefit in the patient population described in the inclusion and exclusion criteria.

Note it links the effect to the patient population as well. So it's not just an endpoint which is capable of showing

change. It links it to clinical relevance. A secondary endpoint or supportive measures related to the primary objectives, and it states that the number of secondary objectives should be limited, and that there should be an explanation of their relative importance and roles in the interpretation of the trial results.

In other words, why are we looking? Is this something that we should be looking at?

There's a great paper by Lubsen that actually talks about combining endpoints into--and he goes through this hierarchy of things that you might want to look at. The first would be all-cause mortality. The next would be nonfatal clinical events. The next would be symptoms of disease, and finally, surrogate endpoints.

The reason he puts them in this order is you can't get to some of the lower things before you pass the ones above it.

For instance, we can't evaluate whether you have resolution of symptoms if you're dead. So you obviously have to count those people because the things higher up here are more important. But we are interested in these multiple aspects of how disease affects patients' lives.

In pneumonia we have death, we have empyema, meningitis or some nonfatal clinical events. So again, extension to another disease is a nonfatal clinical event. We have symptoms like cough, chest pain or shortness of breath, and then we've got surrogate endpoints like cultures, body temperature, white count, respiratory rate, heart rate, blood pressure. Those are all biomarker surrogate variables.

The effect of antimicrobials in severe disease in the past studies was based on all-cause mortality. They really didn't make an attempt to separate out specific mortality, and again, remember from the quote

I showed you, they had the idea that this was all linked together.

So it's also really challenging, if not impossible, for clinicians to determine the cause of death with any certainty.

There's a study out of Germany by Kirch, looked at three decades of autopsy studies, and related it to what clinicians wrote on the death certificate. Now you know what people write on the death certificate most; right? Cardiorespiratory arrest. So that's very informative, telling us that most people who are dead have, their heart stopped, and they're not breathing anymore. But that doesn't really tell us why they actually died.

So what they actually did was they looked at autopsy findings and related to what was written down, and the clinicians were wrong in more than 10 percent of cases. The first thing that was misdiagnosed, pulmonary embolism, and I thought it was interesting that that's what Dr. Wunderink showed this

morning, was the thing that people first thought that person had.

The second most misdiagnosed thing--infection--and it went in both directions. People said the person had infection when they didn't, and they said they didn't have infection when they did, on autopsy.

Doing this will result in misclassification bias, and this is actually Dr. Temple's first article I ever read on the Anturane Reinfarction Trial, about how people misclassified death and what happened when they did that in terms of the analysis.

The other issue is disease-disease interactions are important, and Dr. Mandell pointed out some data that actually shows that. A person may have pneumonia, and what happens is their pneumonia throws them into heart failure, as the Finland data showed. That gives them enough hypoxia that they get an ischemic MI, and then they die.

So we're not just treating the

person's lungs. We're treating the entire person. So it's like a domino effect. If we knock over a row of dominos and they're falling, but I pick up this domino that's already fallen, the rest of those dominos are still going. So it doesn't really make sense to just look at one sliver of the pie and ignore the rest of the pieces.

What are the nonfatal clinical events that happen in pneumonia? Well, this is data from Cecil, you know, Cecil's textbook, who actually looked at untreated people with pneumonia in the past, and you can see that even when people didn't get treated, complications like empyema, meningitis, arthritis, and endocarditis, are actually fairly uncommon, even in untreated people. 6.5 percent of empyema, 1.8 percent meningitis. So even in severely ill people, you're going to have a tough time being able to evaluate these kinds of endpoints because the event rate is just too low.

How can we look at symptoms? We talked a lot about patient-reported outcomes yesterday. Really, what a patient-reported outcome instrument is, is it's an objective measure of subjective phenomena. Think of a thermometer. I don't need a thermometer to tell me whether it's hot or cold outside. But I think a thermometer to tell me whether it's 28 degrees outside or 32 degrees outside. It's a more precise measure.

What's a mercury thermometer measure? It doesn't measure temperature. It measures atmospheric pressure on a bulb of mercury that pushes it up a tube. But we know that that correlates with temperature.

So we just came up with a more precise and accurate way to measure a subjective phenomenon of whether I feel hot or cold. No one questions that. And that's what we're trying to do with PROs.

A PRO is an endpoint measured directly by the patient, with no intermediary.

So it's not an interview by a researcher. It's something that the patient actually fills out. And PRO instruments are the actual tool that we use to measure that patient-reported outcome.

It measures exactly the same thing as the clinician is measuring. We're not asking the person, like: Do you feel great today? We're asking them how's your cough? how's your shortness of breath? and Dave Gilbert showed the top part of the CAP-Sym. It's all the symptoms that we ask people about, except it's asking in a more structured way, so that we're getting the same information from people.

What does that do? It gives us less variability, which gives us smaller sample size, and allows us to actually demonstrate differences with a smaller number of people. Again a good thing.

A number of people came up to me yesterday with the "I don't trust PROs," and

it's something that we're unfamiliar with in infectious diseases, because Laurie Burke has a publication where she shows how many pain NDAs are approved with PROs. And then shows how many ID ones. That would be zero. Actually, I think there's one viral that has something in it related to a herpes drug, but no other NDAs have it for infectious diseases.

How do we put this all together? So we have these number of things in pneumonia that we're very interested in, affecting people's lives. How do we put it all together? Well, it depends on the severity of the disease, as someone pointed out yesterday, and also the event rates for some of these maybe actually be quite low. In mild disease, mortality is low, so it's that event rate, to make sense. Looking at in isolation doesn't make a whole lot of sense.

The other issue is it's not just an issue of sample size. It is clinically relevant. All right. So if I decrease the

mortality from 6.5 percent to 6.45 percent, with a 10 billion sample size, is that even clinically relevant anyway? So it's not just an issue of sample size. It also relates to risk as well. So if I have an itsy-bitsy decrease in mortality but the adverse events increase mortality due to anaphylaxis or liver failure, or whatever, on balance, that's not a good thing, even though I managed to show something on the positive side.

So the issue here, though, is if we evaluate multiple endpoints in isolation, I evaluate mortality and then I look at nonfatal clinical events, and then I look at symptoms, we've got this issue of increasing the rate of false-positive findings by chance alone, which is referred to as the multiplicity problem.

So we can just choose a single endpoint like resolution of symptoms but like I said before, you can't get to resolution of symptoms unless you're alive.

So what you don't want to do is

then exclude deaths from the analysis because that is the ultimate endpoint that we're very worried about, and unfortunately, that is exactly what happens in clinical trials today.

If you say the clinician judges that you're well enough, that you don't need any more therapy, well you have to be alive for that. So if you die early on, you're excluded. So you're called indeterminate if you're dead.

That's about as determinate as you're going to be. So we don't want to eliminate looking at more important outcomes because we're looking at something that's lower down on the hierarchy.

We can evaluate endpoints in combination as part of a composite. So we can focus on a combination of clinically-relevant endpoints like the person is alive with no complications and has resolution of their symptoms.

Pooling them all together will

increase the event rate, and these things are most relevant, as Dr. Fleming pointed out yesterday, when the outcomes are of similar value to patients.

If you would have a composite endpoint and you stick into it, say, your white count going down, well, I have never had a person come to me when they were sick and say, you know, I came to your office today because I really want my white count to go down.

But they do want to feel better and they do want to stay alive. So we can't combine white count with staying alive because those things are very different on a scale of importance to people.

So it also means that if we demonstrate an effect on the composite overall, we can't split it out. So if we demonstrate an effect on death, decreasing nonfatal clinical events, and resolution of symptoms, it'll be driven by resolution of

symptoms.

Can I then say I have a mortality benefit? No, you can't, when you combine those together.

So also, with the evaluation--but we have the urge to do that, and certainly in the voriconazole versus amphotericin empirical therapy trial, people--and all empirical therapy trials, everybody splits out the components of the endpoint because we want to know if there's differences across them.

It's okay to look for consistency of effect to do that, but you really can't make any comments about the individuals ones because you've still got the multiplicity problem.

So if we combine four things together and then split them out again, we've got the issue of multiple comparisons.

So the issue in composite endpoints is failure in any one of the components means you fail overall. So what happens when we add

biomarkers to the mix? So suppose we add somebody's white count to the mix. And actually, I didn't put the graphic in here but this actually happens in some of the Finland data.

You see people who get better symptomatically, and their white count comes down from 20,000 to 12,000. Well, normal white count in our institution's ten thousand. They still have an abnormal white count.

But all the other things have gone away. If we included white count as a part of the composite endpoint, they'd be a failure, even though they felt fine, they were alive, they were doing great.

So depending upon the biomarker, this can actually make it harder to demonstrate an effect, and that certainly happened in other trials, like endocarditis, where you've got to get a blood culture, or the person doesn't get the blood culture, the data's missing, and what do you do about it?

So this is a good way to think about multiple comparisons.

I blindfold you, and I say you're at the bar, and you're tipsy, and you're going to throw this at the dart board, and you have a one in twenty chance of hitting that bullseye just by absolutely accident. But then I add a few more bullseyes to the dart board. Now you've got a higher chance of hitting one of these bullseyes by absolute complete happenstance.

So what can I do about this? So this is the multiplicity problem and this is a slide I borrowed from Bob O'Neill. What happens when you do one comparison? You have a 5 percent chance, if these things are completely unrelated, of making a mistake.

If you get up to 10 percent, you've got over a 20 percent chance of making a mistake by complete accident, almost to the point where you're surprised when you don't find some kind of a difference.

There's articles written on this that say the average number of subgroup analyses and secondary analysis is around 18 for clinical trial and some go up as high as forty-five.

So imagine how many comparisons you're talking about. What can we do about that, to control for this? Well, the one thing we can do is shrink the bullseyes, so that all the bullseyes, now added together, add up to the size of the one bullseye before. That's called adjusting the Type 1 error. It means you split up your p value amongst all these things.

But when you do it that way, what happens is your sample size goes up and it doesn't go up linearly. So if I want to evaluate 16 endpoints and I split up my p value among 16 endpoints, I have an enormously large trial I've got to deal with.

But how about I do this? How about I blindfold you and I say I'm going to show

you one dart board and you're going to throw at that one. Then I'm going to put up the second one, if you hit the bullseye and you get lucky, I'm going to put up another one, and you have to hit that one.

And if you hit that, then I'm going to put up a third one and you have to hit that one.

This is called serial testing or hierarchical testing, or a gatekeeper approach. Each one of these, you individually have a one in 20 chance of being able to hit the bullseye.

So by doing that, you can actually answer multiple questions without increasing your sample size. But that requires putting these things in some kind of logical order.

You wouldn't want to put white count first and then death second. You'd want to put the most important thing first, because the problem with this way of looking at things is if you lose, and you don't hit the bullseye

you don't get to go on to the next dart board.

So that's why you have to put this hierarchy in a really rational way, and think about how you lay those things out.

So what are some potential approaches in CAP? I'm kind a "jumping the gun" because we haven't heard about this yet. But it appears that looking at this data, that the endpoint for non-inferiority trials, and there does appear to be an effect in severe disease with people with pneumonia, but the endpoint for these studies was all-cause mortality.

And that's really the only basis that I could find, looking through these trials, for an endpoint. We don't know about these other endpoints that we like to look at. But that doesn't mean that we shouldn't look at them.

We can test these other hypotheses, perhaps in a hierarchical approach, and then we don't need to adjust for Type 1 error, but

we need to show superiority on those endpoints, other than all-cause mortality.

Does that mean your drug wouldn't be approved? No. If you show non-inferiority on all-cause mortality, that would be the basis for approval, but we could answer some really rational questions with the other endpoints, like, How long do I need to give people therapy on a time-to-event analysis? Does my more potent drug make people get better faster than the other drug? And if it doesn't, you haven't lost anything because the approval is based on the effect on non-inferiority on all-cause mortality.

So we can answer really relevant clinical questions and get a drug out there for approval, for people to use, all at the same time.

What's really not clear here, though, is how biomarkers add anything in this evaluation of response. So we heard yesterday that Dave Gilbert showed chest x-rays just

aren't responsive to change. The micro data just isn't there. Over 90 percent of people in these trials don't have follow-up microbiological data cause they don't have it to give.

They're not coughing anything up. They're better. So there's really nothing to analyze here, and we don't know how other laboratory measures actually work out.

Why is it that clinicians think that heart rate, blood pressure, and all those things are clinical measures?, cause that's what I use every day when I go in and I evaluate the patient. But what am I trying to do in that setting?

I examine the patient Monday because I'm trying to figure out how they're going to be on Tuesday, Wednesday and Thursday. That's clinical practice. In a clinical trial, I measure how they are on Thursday. That's what I want to know, not trying to guess what's going to happen in four

days.

So we actually just want to, in clinical trials, to get out just a little further in this setting and actually measure what's happening.

Why do we, in HIV, always use a biomarker? Because that taken out further means I'd have to follow the patient for months to years. This is not a disease where you have to follow people for months to years. We can actually find out what's going on.

Just to sort of step back to yesterday, it's really difficult to find anything related to mortality in mild to moderate disease. So you could show superiority in either dose response or superiority to another agent, or a placebo-controlled trial, on a composite endpoint of mortality, nonfatal clinical events, and time to resolution of symptoms.

What will drive that endpoint? It will be time to resolution of symptoms,

because that's what most people would have. But you shouldn't ignore the other more important things as well. Again, we talked about how PROs can objectively measure subjective phenomena, and again, those PROs can be used in severe disease, not as the primary endpoint, and in an ICU patient on a ventilator, obviously, that's not going to help you at all, in that particular setting.

But they can help us in terms of the ward patient, time to getting better, actually help us to make some decisions about duration of therapy and other things. And again we talked about how there have been PROs in community-acquired pneumonia that have been evaluated before.

The issue of multiple comparisons also applies to subgroup analyses, and the issue here is that the subgroup analyses really apply to this idea of looking at outcomes by organism.

The 1992 points-to-consider

document on antimicrobial development suggests that you should have a number of isolates, at least ten per organism.

So that gives us an ability to look across a breadth of organisms, and try to get an idea that the drug is similarly effective in all those areas. But when you have very small sample size, you can be easily misled.

For instance, in intra-abdominal infections, tigecycline had a success rate, I think it was like three out of three for Pseudomonas in intra-abdominal infections. The drug doesn't have any in vitro activity against Pseudomonas.

So you can see things by just chance, when there's very small numbers, and when they're not sufficiently powered to make claims regarding superiority of one drug to another. So what would happen if you wanted to develop a drug, and it's out there, there's loads of resistance like MRSA, like macrolide-resistant Strep pneumo--how could we actually

do a trial that looks like this?

Well, we already started to address this. What you'd want to do was actually power that you could look at disease due to resistant pathogens, and show superiority in that group, and then look at another complementary group of the people who don't have resistant infections.

What happens, though, is that people look at the subset of people who have resistant pathogens, and then look at the overall trial results and try to demonstrate non-inferiority there.

The theory here is we're saying our drug has benefit in a predefined subgroup of people, who we expect our drug to be better, because the older drug isn't working so well anymore.

What we then want to know is, we don't want there to be no effect at all in the people that don't have that pathogen. So we want to evaluate that the people who don't

have that pathogen aren't being harmed. So we want to say--we're asking two questions here. So it's essentially like two trials in one. We're saying is there a similar effect in the people who don't have that pathogen? and are we better in the setting where we think we're going to be better? So if you look back at the overall trial results only, and you have a spectacular result in the resistant pathogen group, and the resistant pathogen group is big enough, it can actually drive the overall results and be hiding the fact that you're actually worse than nothing in the group that doesn't have a resistant pathogen.

So, for instance, suppose I took vancomycin and I study it in people with pneumonia. 99 percent of the people have MRSA, and I use it, and I show superiority to whatever. Let's just say it's placebo; but we would never do this. Let's say we go back to the 1950's and we find a trial of vancomycin versus placebo for MRSA.

But then I look at a group with Gram-negative organisms, and there's five of those people in the trial, and they look like they did okay.

Would we use vancomycin to treat Gram-negatives? No. It's only because the overall result where we're having an effect is driving the whole thing.

So we need to power both of these pieces individually, to be able to make any kind of statement about where we are in resistant pathogens.

Now for some things like MRSA, it's becoming common enough that you could probably do something like this. Things like vancomycin-resistant Staph aureus are just not common enough, at this point, to be able to do a kind of trial like this.

So this really relates to things that we can actually study.

So to finish up, then, appropriate selection of endpoints should include an

evaluation of what to measure, which is content validity--and I want to reiterate the point that things like content validity, construct validity, etcetera, they apply to any endpoint. It's not just something about PROs. So we need to know what to measure, how to measure it and when to measure it, and how much change in that endpoint actually makes a difference to patients as well.

We also need to evaluate clinically relevant outcomes in a timeframe that's relevant to the natural history of the disease and that'll provide us better information in a more efficient manner, but we also need to take into account the issue of false-positive results with multiple testing and consider various approaches like this hierarchical or serial testing approach, which would allow us to answer multiple questions in the same trial without having to increase the sample size of the trial dramatically.

So I'll stop at that point. Thanks

much.

DR. FLEMING: Thank you, John, for what was certainly a far-reaching summary of those issues. Dave, I think we're going to go on to the next session here without, next lecture, without Q&A's; is that right? To stay on schedule?

DR. GILBERT: Well, that was a great intro to our next presentation. Now we need to "drill down" to clinical and microbiologic endpoints, and Dan Musher from Baylor is here to discuss this topic.

Dan.

DR. MUSHER: Thank you, David, and thank you to the group for inviting me, and of course I wish that--the problem with speaking from slides is I prepared the slides a few days ago but I know a whole lot more now than I did then because of all the discussion I've heard.

So I'll have to modify them as we go along. Philosophical problems, and this

has already been dealt with at great length, the natural history of the disease that we're talking about influences our interpretation of cure or failure because of varying proportion of response spontaneously, and that proportion is a big subject of discussion.

It varies with etiologic agent and it varies with the severity of the disease. If you've got a mild case of pneumococcal pneumonia, however you define that, you're going to get over it, and if you've got a serious case, you're not, unless you get antibiotic therapy.

And mycoplasma I think gradually resolves over a period of time anyway, but its resolution can be hastened by antimicrobial therapy.

Generally, there's a very high success rate of existing therapies for existing pathogens, and that could change with the emergence of a new pathogenic organism that caused disease, or with newly-resistant

organisms.

And we've already heard a lot about this whole problem of empiricism. In many cases, we don't know what infection we're treating, so it makes it awfully difficult to--you know, you lump everybody together and you've got some people with viral pneumonia, and some people with Mycoplasma pneumonia, people with pneumococcal pneumonia, and lump them all together, and they get better, and well, how do you really know?

So we might not be so certain that our drug is producing a cure.

We certainly should be able to develop criteria to recognize therapeutic failure.

Now what constitutes a clinical failure of treatment for pneumonia? And I'm going to have slides that discuss each of these things as we go along. So I'm just going to go ahead to the next slide.

Death would be a good one. So let

me just point out now, clinical failure of treatment for pneumonia, because that's the focus. So there are the famous Austrian and Gold figure, and I don't think anybody disputes that, suggests that death within that first 72 hours is a result of cytokine storm. I mean, it's just going to happen, whether you've got an effective antibiotic or not.

So if you're trying to determine whether your drug treatment is correct, you might ought to exclude death within the first 72 hours for your analysis. I'm just pointing that out. That subject hasn't come up but you might want to consider that.

And you know, that's after 10 or 14 days. I mean, those of us who take care of patients know that the deaths are a result of all those comorbidities. The lungs fill up and then you give them a diuretic, and then the kidneys start to fail and then you give them fluids, and then the pulmonary edema gets worse, and eventually, one thing leads to

another, and unfortunately, in American society today, the morbid obesity kicks in and all of a sudden you've got a dead patient.

So it seems to me that the mortality, somewhere between three days and ten days, probably is the best indicator of how effective your antimicrobial therapy is, and that's what I'm pointing out in this particular slide.

What constitutes clinical failure? New or persistent or recurrent bacteremia by a causative organism while the patient's on therapy.

So we've seen that in the studies of Staph aureus bacteremia. We've seen that vancomycin, on vancomycin, bacteremia persists for days, and that does not happen when you treat a methicillin-susceptible Staph aureus with nafcillin.

And the question is, well, is that just some kind of epi phenomenon, or does that relate to how well or how badly they respond?

And the data, in the case of Staph aureus bacteremia, do really show that if you've got bacteremia going on for several more days, that is associated with a higher rate of complications and a higher rate of mortality.

The thing is in pneumonia, it's very uncommon. In community-acquired pneumonia it's a rare occurrence. You can have Gram-negative rod pneumonia. In severely immunocompromised patients, repeated bouts of COPD on many courses of antibiotics and steroids, and obviously if bacteremia recurs it's a failure.

But the percentage in which that will be seen is way too small to be useful, and of course that doesn't even include all the people who do badly, who don't have bacterial pneumonia the first place.

How about complications such as necrotic lung, empyema, infection at a remote site? We've already seen a table showing that

even in the pre-antibiotic era this was uncommon. These are uncommon. Now I think if you add them all up in my hospital, they add up to five or six or seven percent of all the proven pneumococcal pneumonias, and many of those have been set in motion before your treatment was begun.

So therefore, the appearance of these, the symptoms become manifest on the third or fourth or eighth day of treatment, but the thing might have been "cooking" prior to therapy anyway.

However, it's so uncommon, because they occur in such a small percentage of cases, it's just going to be difficult to measure that. It can be hard to know what to do with the information.

Delayed defervescence. This one was used historically, and it wasn't in comparative trials because they didn't have comparative trials.

I do think that that is a perfectly

fair measure of how effective an antimicrobial agent is. I think the patients, if you take a number of patients on one drug, and the number of patients on another, and if the rate of defervescence is less, then the antibacterial effect is also less.

I can't exactly prove that but it sure seems to fit with whatever principles I think I understand in infectious diseases.

However, even that's compounded for the following reasons. If a patient's on his way to a cure, does a day or two of temperature above, between a 100 and 100.5, does it make a difference or not? Well, I don't know.

And is the defervescence due to some other property of the antimicrobial agent? You've got this whole dispute over whether there's an anti-inflammatory component to the macrolides.

Obviously, failure to defervesce is consistent with clinical failure, although

other causes are possible. So I'm still dealing with this clinical failure of treatment.

Other possible considerations. I think days in an ICU are a pretty good indication that something isn't right, and has been pointed out by several people, most recently John, you always have to factor in the people who have--they're only in the ICU for 18 hours because they're dead. So you have to have some--and I'm not anywhere near clever enough with statistics to know how to handle that--but I know you can't ignore it.

How many days a patient remains intubated. Now, in part, that's determined by the underlying state of the pulmonary disease, but I think that's also--I think that if we have a more effective antimicrobial agent, they're going to have--the other things will even out in your randomization. Then you'll have fewer days of intubation in the group that's treated with a better antibiotic than

if there's a less good antibiotic.

Days of IV therapy, if there are totally blind protocols and you have an option to switch from intravenous to an oral therapy, and it is absolutely blinded, then, in the clinician's judgment, my patient is well enough to switch over from an IV to oral therapy, and maybe that's valid. The point was made yesterday, I thought it was a good one, I hadn't thought about it, that everybody who puts a patient on protocol knows the patient's getting something, and that does really inform the thinking about the cases.

Total days in the hospital is really too dependent on comorbidities. In preparation for this meeting, I reread, very carefully, the papers on the time-to-clinical stability and the symptom questionnaire, and I've got to tell you guys, I think these things are really very valid. I think there's a lot of good reason in them.

And I think if you create graphs

from the data and you show the slope of curves, and statisticians can do that, I can't but they can, and if you have a more effective antibiotic, there will be increased rapidity of time to clinical stability using the criteria that are set out, and I think this is a very nice--it's a very nice set of criteria.

And the same thing with the symptom questionnaire. It's exactly the same. I've had, believe it or not, my two grown daughters have had pneumonia this past month. One of them, clinically, was a perfect example of the pneumococcal pneumonia, one was a perfect example of a Mycoplasma pneumonia, and they got well over a varying period of time.

And you take enough patients, and you average them out, and you can just tell. Are they getting well? Is it a steady improvement? Is there a relapse? You can use a white blood cell count.

John, one of them, the one who's a "hot shot" internist--she's a med peds person-

-she started--her temperature went back up a little tiny bit. Daddy, shall I get a white blood cell count? I think you better, because I think if it's going back up we've got to look for a complication. So the white blood count was 6400 down from 28,000, and the fever went away.

There is a place, that's why we clinicians use them. I don't think that these studies would exist if we didn't find some use for them, clinically. So I think that the time to clinical stability and the symptom questionnaire really do make a difference and I think there are ways to evaluate the slopes on those things.

Microbiologic cure. This is going to be very simple. I want to comment about microbiologic diagnosis and I want to add to Tim Murphy's comments yesterday. I did this study and published it in Clinical Infection Diseases, and if you haven't seen it, I'm not tooting my own horn, but have a look at it.

People talk about the Gram stain and culture of sputum not being reliable. It's only 40 percent of the time do you get the diagnosis.

So I took 100 patients at my hospital who had pneumococcal--they had pneumonia with pneumococcus in the bloodstream. I think we would all accept that as a definition of pneumococcal pneumonia.

And only 70 percent of them--oh. Overall, the Gram stain and the culture showed pneumococcus in about 40 percent. That's exactly right. That's what everybody says.

Well, now look at this. Denominator was a 100, only 40 percent give an answer, but only 70 percent of them had a sputum sent in the first place. Well, you can't evaluate the validity of a test if it wasn't even done, and then the laboratory rejected the sample in another 15 percent. They said it's not a valid sample. So there's no analysis done.

You can't include that in the denominator. So then you only have 55 percent of them, and actually that 55 percent, now your sensitivity was something like 70 percent, 75 percent.

And then it's a paper with two figures in it, is all it is. And then the other one, I looked at how many hours they'd gotten antibiotics, and the ones who had gotten no antibiotics at all, the sensitivity of the sputum, Gram stain, in culture, was about 85, 90 percent. That's pretty good as tests go in this world, and, actually, the sensitivity remained pretty good up to 12 or 18 hours by culture, not by Gram stain, and those of us who've done this know the bacteria go away pretty fast by Gram stain but you can still culture them after 18 hours of antibiotics, and after 24 hours, which was a bunch of the sputum samples submitted from my hospital, after 24 hours you can hardly find the organism anymore.

So you had a lot of problems with microbiologic diagnosis, and if you can't make the diagnosis in most of your patients, you can't evaluate the cure. The notion of microbiologic cure--guys, just forget it, drop it from--my recommendation--if you ask me my recommendation, it never belonged, people were getting invalid samples or sending saliva, they would swab on these soft tissue infection studies--they'd swab the skin and send something to the lab because the drug company wants that little box filled in. It's craziness. It's craziness.

And it's compounded further by the fact that you've got colonization, as Tim Murphy has shown very nicely, and you'd have to fingerprint the organism. If somebody started out with a Haemophilus pneumonia, five or six days later, still has Haemophilus but it's totally cured, that's because that person is a long-time chronic bronchitic colonized with Haemophilus.

But even if you thought the Haemophilus isolate, five or six days later, was meaningful, you've got to fingerprint it and fingerprint the original one and show that it's same organism, which 50 percent of the time it is and 50 percent of the time it's not. It's very complicated. Just forget it. The patients get well, and if the patients don't get well, then you work them up as if they've got an infectious disease problem, and you go ahead and look for a pathogen.

So anyway, I just don't know what to do with a bacteriologic cure. Here, by the way, someone commented on this earlier, this was the actual quote from Dr. Finland. "Pneumococci were eliminated from the sputum in 50 percent by 48 hours, some persisted for five days or more--this is a quote--probably related to low doses of penicillin and once larger doses were used, the clearance was more rapid."

This J. B. Amberson lecture's a

beautiful reference, in case you--Dr. Finland never wrote things short. This is about a 25-30 page article but it's got a lot of data in it.

So if it's difficult to establish the diagnosis, it's more difficult to establish the efficacy of treatment, and that's all I wanted to say.

So summary conclusions, evaluating clinical and microbiological responses during treatment of what is called community-acquired pneumonia, and my big sermon that I give the house staff, I can't even give this crowd, I don't have time. Community-acquired is not the name of an organism. It's where they got it. Pneumonia is caused by specific organisms and we should be looking for specific organisms, and aiming specific antimicrobial therapy, and if you take care of patients, it's a totally different perspective, guys. Totally different.

Anyway, I think that the symptom

questionnaire is useful, and the time of defervescence, the time of clinical stability are useful, if you want to evaluate, if you've got an effective antimicrobial therapy, and if you want to compare A to B, I think you can use those things.

You can look at mortality between 72 hours and 10 days, and except in patients who are hospitalized, who are very sick, it's not going to be a high enough number you're going to get much use out of, certainly not--well, we already discussed that. Length of stay in the ICU, the days of intubation, I think does--since I believe that the underlying badness of the lung disease will average out in your randomization, then effectiveness of your antimicrobial therapy will be, will affect the length of stay in the ICU, and will affect the days of intubation.

Development of a complication on treatment will be very uncommon but obviously it's an indicator of the effectiveness of your

therapy. Emergence of resistant bacterium--you've got to prove it's the same organism and then you really need to show that it's associated with a clinical failure.

Otherwise, it's just bacteriology, and of course persistent bacteremia. So I don't know if that's helpful or not helpful. I had a good time thinking about it. Thanks very much.

DR. GILBERT: Thank you, Daniel. We're going to take all the questions and comments at the end, and I'm supposed to hit escape. Do I hit escape again? Maybe.

So next, in order to--how come I'm not getting page two? Here we go. Thank you. I didn't even touch it. That's the next page, page two. Yes. Boucher. Top one.

Okay. In order to interpret endpoints, we want accuracy, we want to reduce bias. Of course a major way to reduce bias is to ensure appropriate blinding and this is definitely not "the blind leading the blind."

Helen Boucher has a great deal of experience and insight in this area. She's from the Division of Infectious Disease at Tufts-New England, and Helen, can you help us out here.

DR. BOUCHER: Thanks very much, Dave, and Drs. Fleming and Cox, for inviting me.

I confess, that when I saw the title of this topic, Is it possible to blind a trial of CAP? I wondered really what I was supposed to address, and I'll sort of share a little of how I got to where I got and try to leave a couple of messages that I think are relevant for us, thinking about blinding trials, especially in our severely ill patients. My conflicts, or potential conflicts are listed here, and, you know, to start, I think the answer is yes and no.

We always hear about blinding trials, but I think to really do it well, what I learned in this exercise is that that's

actually pretty difficult. It may not be the best answer in trials of our seriously ill patients.

if one just did a PubMed search on blind and community-acquired antibiotic, you'd get 139 articles. If you did pneumonia, you'd get 576. There's a ton of the use of blind and titles of our trials.

So to try to make my life easier, I said, well, let's look at what's been approved by the FDA since 1998, and the subgroup who studied community-acquired pneumonia. These drugs weren't necessarily approved for community-acquired pneumonia but there are published studies in community-acquired pneumonia.

And thanks to Brad for sharing some of his data on this.

The message here is pretty impressive; right? I mean, if you look at the far column, the yeses way outnumber the noes, and the noes were early in studies published,

you know, the two drugs approved in the 2000 range. So what's the question?

Well, when you go and read the trials, it gets a little more interesting. So I picked three, no particular reason, no statistics, just looked at the ertapenem trial, the gemifloxacin versus trovafloxacin and gatifloxacin versus amox/clav. They all are double-blind trials but the only place you see any discussion is in the title, the abstract, and the first sentence of the study design.

This was a double-blind, and in one case, double-blind double-dummy trial. There's no description of what they did, the groups did, or if they assessed blinding. And when I went back to the consort guidelines, there's actually a whole page in the consort guidelines now that tells you, when you're reporting a trial, steps for reporting the adequacy of blinding.

So one trial I found that's

actually an order trial, is a trial from Fink et al, that was published in the AAC in 1994. This was an early trial of imipenem versus ciprofloxacin, and these authors actually said in their intro that our goal was to achieve a, quote, better blind, to use a double blind, and we conducted and analyzed this study under fully blind conditions.

They include, when they describe the treatment, that the pharmacist was unblinded, everybody else was. They talk about the actual dummy infusion and they make it very clear that their decisions about how to handle these dreaded premature discontinuations, that they were made prior to unblinding, all things about evaluability were made prior to unblinding, and they interestingly tell us about how many people had to get the placebo for the metronidazole, which was sort of interesting, in and of itself.

The assessed cause of death, which

comes back to some discussions we've had this morning, prior to breaking the blind, and they did their analysis before unblinding.

And the reason I mention this is that I know from industry and FDA colleagues that you all spend a huge amount of time on this and there are whole divisions of companies that work on making the dummy pills and capsules, and all that.

But I think we and academics don't give it enough attention, and when we review articles and stuff, a lot of us are negligent because we don't go back and ask the authors to tell us more about what they did.

So does it all matter? You know, should we care? I found a very interesting study, cohort study that was from the Cochrane review. They took 200 randomized trials published in 2001. 78, over three-quarters described double-blinded trials. 56 percent, over half, didn't tell anything about who was blinded. A quarter didn't tell any more than

that it was a double-blind trial and 2 percent explicitly talked about the patients, the providers and the data collectors, and how they were blind.

And then what I think is really the biggest take-home is here, that they asked everybody, you know, what does double blind mean? and they got 15 different definitions, and everybody thought that their definition was it.

So I think that the message from this is that our interpretation of blind is different, the reporting is certainly inconsistent, and that they also brought up the notion that the assessment of blinding was lacking, and we'll come back to that a little bit later.

So in terms of blinding, we've heard a lot about, you know, who should be blinded? Our patients, the investigators, the people doing the outcome assessments. Sometimes that is or isn't the investigator.

Data analysts. I'm not going to spend a lot of time on this but it might be worthy of discussion, about some of the details of Data Safety Monitoring Boards, or DMCs, which we got into yesterday.

We haven't mentioned anything about review or adjudication committees, and that may or may not be so relevant in this area.

And then what exactly should we blind? We all would agree that the study drug should be blind. But what about the microbiology? Do we need to know, and what do we need to know? When, I think is important. The outcome assessments of both efficacy and safety we'll comment on, and then I'll spend most of the time on the challenges because I believe in our seriously ill patients, there are a number of challenges that are very real for us to address.

So before I delved into this, I learned something very important about definitions here, and that is that when we

read about trials, a lot of what we read doesn't distinguish between allocation concealment and blinding. So I just wanted to go over that, briefly.

Allocation concealment is really what comes first, before the patient gets randomized, and that's keeping everybody in the dark about who's going to get what, and that prevents selection bias, and so that keeps that sequence, the list of what A and B is, totally away from everybody before and until that assignment is made.

And that can always be done, that can be done in open label trials, and I learned from Dr. Fleming, a long time ago, that has to be done well, and everybody has to know the details of that, especially in an open label trial.

Blinding or masking, and if you look at older literature, a lot of the articles talk about masking, but nowadays it seems that blinding is the term. That's

keeping us in the dark about what intervention assignment was. So is keeping our patients, us, and the outcome assessors blind.

And this prevents ascertainment bias and this protects that sequence after the allocation, and this is where I think it's harder, especially in our ill patients, to keep everyone in the dark.

So the potential benefits a lot of us have gone over, but for our patient participants, they are less likely to have biased responses to the drugs. They're more likely to adhere, and that's really important in the non-inferiority setting where losing people is so pricey. They are less likely to ask for extra therapy and less likely to leave, or get into that lost to follow-up category.

For the investigator side, we're less likely to transfer our preconceived notions about a drug to our patients. We're also less likely to selectively do other

things--do more diagnostic tests, give other therapies. It's also been proven that we're less likely to address the dose, which I thought was kind of interesting. Less likely to differentially withdraw patients. You're less likely to have your colleague come in on the weekend and say I'm just not comfortable that they're on the experimental drug, because I know they are.

Less likely to differentially encourage or discourage your patient to stay in the trial. For the assessors, I think everyone would agree that they're less likely to take their biases in making outcome assessments.

So the level of blinding is shown here, and most of the trials that we are focusing on are double-blind, and that means that the patient, the physician-investigator, and the assessor, who may or may not be that physician-investigator, are blind.

Like we said earlier, the

terminology is confusing and it's probably more important to tell what we did when we described the trial than rely on these terms.

So let's talk about some of the challenges now and I'll start with some feasibility issues. So matching is a notion that the capsule, the tablet, the IV bag, matches in the two groups. And this can be a really "big deal" and it frequently involves a new formulation for the study. I was involved in some tries to blind amphotericin, ten years ago, you know, with shrouded bags and dummy tubing, and it can be a very "big deal." And it also goes to the extent of masking the color, the odor, the taste. And I know our industry colleagues are very familiar with this.

Some of the other details are that the containers have to look the same, the codes have to be the same, and it is important that we talk about the potential inadequacies of matching in our publications.

So, you know, the formulation I mentioned, this notion of enclosing the dummy and the active agent in identical capsules is very attractive, but that's actually often not so feasible because of expense, time, and making the capsules so big that the patient can't take it.

The double dummy I think is pretty intuitive to many of us.

In the IV medication, I think this is where this becomes a "big deal" and where we encounter one of the biggest challenges that we probably should discuss.

When we have a dummy or a placebo IV, you introduce volume load, different frequencies of administration, the need to use the precious IV access, my patient in the ICU who has a triple lumen, that's getting TPN and blood transfusions. Now I need to take it for two more hours. Even on the floor, that can be an issue.

Drugs that require therapeutic drug

monitoring are particularly challenging. So vancomycin is the one that's probably "near and dear" to a lot of us. Sometimes we change not just the dose but the interval for vancomycin. We might go from every twelve to every eighteen.

So is it possible to prespecify in a protocol, that we would just adjust dose and never interval, so that we could keep things blind?

What about that unblinded pharmacist? How feasible and practical is that, and necessary, because it introduces a huge expense to always have that person on board?

And then one I think that a lot of us forget about is the labs. So you've got to keep the labs secret too. The vanco level. I can't know the vanco level, because if I know the vanco level I'm going to think I have to do something. So keeping that from the investigator and the study team is

operationally often quite a challenge.

Then George Talbot yesterday spoke about concomitant antibiotics and the challenges they bring, and I think the narrow spectrum agent question is difficult because, you know, you want to add something that covers what you need to cover, without interfering for my ability to know the drug I'm studying works.

So we've seen this in hospital-acquired pneumonia studies and some skin studies. The whole question of aztreonam for Gram-negatives, are we comfortable with that? At my hospital, aztreonam is a lousy drug. So I'm nervous about that. Some people have said you can use some Pip-Tazo for a little while. Well, how long? We've heard about the challenges of even one dose of antibiotics. So if your drug and broad spectrum overlap, is that going to, you know, cause grief in interpreting your study.

What about geography? We haven't

heard that much about it. You know, the comparators we talked about are different. Standards of care are different. So your adjunctive therapies are different.

Is it ever going to get to the point where we have to study different comparators and use different techniques in different parts of the world? And then that's where the resistant issue comes in. Our local epidemiology really does vary a lot, not only in the United States, but if you go to Southern Europe, and different places, the rates of resistant organisms are much higher.

And I think we have to think about this, both in how we might conduct and report our trials, but then how generalizable they are at the end of the day, and if we're going to achieve what we want to achieve.

So a little more about microbiology. We've heard a lot about this. Obviously, we're all going to be capturing it. We're going to be, you know, doing the

aggressive measures that we talked about this morning. But do we have to know about it?

So are there some circumstances where we have to know what the "bug" is or what it's susceptible to, now, in real-time?

So the strep pneumo issue has been raised, but are there circumstances where one might need to know if it's a resistant strep pneumo? The community-acquired MRSA, we haven't really discussed. But that's a big problem. We've lost some patients, recently, and a lot of my colleagues, and maybe me would think--I need to know if this patient has community-associated MRSA. I don't know if I'm comfortable, that's an issue, and then resistant Gram-negative rods.

We do have people coming in from the community with KPR-producing organisms in this country. So that's another thing we probably want to discuss.

And then, if we're going to find out, should we do anything about it. Do these

tentative break points fit in? You know, should we be making therapeutic decisions based on MIC type data during the conduct of a study?

And sinusitis has been addressed, I think it's been alluded to at this meeting, and the recently released guidance says that when we do sampling, the investigator should know about it. Can we extrapolate that to pneumonia or not? I think that's a question people want, we should discuss.

What about our outcome assessments? A lot has been said about this in terms of the relative ease of handling hard endpoints like death, because they're less biased. And I think emphasizing that blinded assessors are advisable even in open-label trials is important, and something to keep in mind if we think that an open-label design is important here.

In terms of safety, I think we don't always consider, but knowing the drug

someone's on will also influence the way we respond to certain adverse events, and I think the fungal experience with the amphotericin's a good example.

We would tolerate a lot of adverse events in our patients getting ampho and press on. We'd keep going when the creatinine went up, we'd just draw up the Demerol and keep going. So I think we wouldn't necessarily see people reporting that as an adverse event. So that's another reason, where blinding can help us in terms of safety.

Now the ethics, probably the most difficult to grapple with. But I think our job is to, you know, make our patients, our colleagues, comfortable that nothing bad is going to happen in either group, and that we have to--you know--everybody has to be convinced of that, that both therapies are acceptable, and in this sick patient population I think that is maybe not always so easy to do. The notion of delayed or rescue

therapy has been brought up.

I think that my position, after seeing what we've discussed and what I've researched, is that it's not an option in our seriously ill hospitalized patients.

And what about when we do rescue patients, when they do fail? At that time, do I need to know what the patient was on to properly rescue him or her? That's also an issue.

So this whole notion of unblinding. There are things that are unintentional we can do to unblind a patient. That's if the drugs are labeled wrong or something happens in a logistical issue.

The laboratory thing is something I've seen, where labs accidentally come back because the central lab accidentally sent them back to the site.

And I think when we decide to intentionally unblind, or intentionally withdraw a patient, it's important to

emphasize that we should always just try to stop the patient rather than unblinding, and if we're going to have criteria for unblinding, we should make them very, very clear at the beginning. That it's going to be patient safety. Only the people who need to know need to know when that unblinding is done. Only the patient/treating physician, not everybody else, and perhaps only to deal with unanticipated safety issues.

But the real take-home here is to have those strict criteria for breaking your blind before you ever start.

The conduct of the study is really important with blinding, and I think the decision to continue a patient or switch to alternative therapy really helps. It's much better when you have a blinded study. You have less of this differential loss that we've talked about a lot, and that's important in non-inferiority trials.

This whole notion of stopping due

to lack of efficacy, it's one that we've struggled with in a lot of areas, but is it feasible to prospectively define reasons to withdraw?

You know, Dr. Musher brought up the Staph bacteremia. Is seven days of persistent bacteremia, is that a reason to discontinue somebody? I would say maybe, because if they have seven days of bacteremia and they're better--otherwise--and I have nowhere to go, I can't take out their dialysis catheter. They might be able to stay.

Someone with two days of bacteremia, who's in septic shock, on their head, they might have to be withdrawn due to lack of efficacy. So while it sounds good to prospectively define these reasons, I think that's a big challenge.

To the extent that we can do it, it's certainly better. And then I come back to that question of when you do pull somebody out of a trial for lack of efficacy, do you

have to know what they were on, and what's the impact of that?

And finally, this notion of assessment of blindness. This is in the CONSORT statement and sort of throughout all the textbooks, that one should undergo some kind of exercise to try to know how well the blind was maintained, like ask the patients or the investigator to guess what group they were in, and the guesses should be random.

And if they're not random, that may tell you something about the degree to which the blinding was successful.

One author went so far as to say you should actually do that, and then measure it in each--for the whole trial and then in each site, and that's sort of a big burden. It leads to some interesting potential implications, I think.

So what about if blinding is impossible, if we decide that in our sick patients we just--it's not the best way?

We've established that laying out a rationale for that is very important. Tell how we minimize bias other ways. Allocation gets--I mean is key. Trying to have our clinical assessments made by others, that can be blinded, and leaning on endpoints that are harder, like death, and potentially micro endpoints, when you have things like bacteremia.

So to come back to where we started, I think blinding is possible in these trials. But there's a cost, and the cost is not only in terms of doability and execution of the trials, but potentially to our patients. When you get into giving people big sodium loads, who are already very sick, you know, that's not always a trivial thing, and I think that's worth some discussion.

If we're going to not go the blinded route, then we have to be even better in a lot of the other aspects of our trial conduct and design.

And then I really don't know what the answer is about blinding of microbiology data in our seriously ill CAP patients, and we'll probably want to discuss that some more.

So with that, I thank you for your attention.

DR. FLEMING: Thanks, Helen.

Now I'd like to ask Mary Singer to come to the podium. Mary will be talking to us--she's a medical officer in the Division of Special Pathogen and Transplant Products, and she'll be talking to us about the work that she and others in FDA have been involved in, in looking at the historical data, to try and understand treatment effect in community-acquired pneumonia.

Mary.

DR. SINGER: Good morning, everybody. My title is a little bit different than what's in the agenda. I'm going to focus mainly on the treatment effect of the antibacterial drugs in community-acquired

pneumonia from both a historical and regulatory perspective, and I'd like to mention that I have no disclosures.

Today, I'll discuss the problem with non-inferiority trials for community-acquired pneumonia, in brief, the approach to, our approach to estimation of an antibacterial drug, treatment effect in CAP, the estimates of the treatment effect, limitations of the data, and then present the issues for further discussion.

First, I wanted to put this in some perspective. I wanted to review briefly, what Dr. Higgins talked about yesterday, what we've seen in recent CAP studies.

So far, about 30 antibacterial drugs have been approved for CAP. The recent studies have all been based on non-inferiority trials. Most have been in patients with mild to moderate CAP, treated in the outpatient setting with oral drugs.

Pneumococcal pneumonia has been

documented in 5 to 20 percent of patients in oral drug studies, and up to 20 percent of patients, in hospitalized patients, with studies of initial IV therapy.

Bacteremia, documented in 0 to 6 percent of patients in oral drug studies, and 8 to 10 percent in IV drug studies. 4 to 9 percent of the latter was pneumococcal bacteremia.

Efficacy rates were high, across the board, using clinical response as an endpoint. Mortality rates were very low, in general. Less than 1 percent of patients died in the oral drug studies, 2 to 4 percent in the IV drug studies.

So as clinicians, many of us would feel very uncomfortable, not treating a patient even with mild pneumonia. So what is the problem here?

In non-inferiority trials--and I'm just going to go over this briefly, because this again is a review of what we talked about

yesterday--we're asking the question, How much less effective is the test drug than the active control drug?

The efficacy of the test drug must fall within the bounds of a pre-specified non-inferiority margin relative to that of control drug. This assumes that we know the treatment effect. So we know by how much the active control is more effective than placebo for treatment of the disease. So this is called M1, or the treatment effect.

So if we know the treatment effect, we can choose then a clinically acceptable non-inferiority margin, or M2. That's always less than or equal to M1.

The problem lies here. We don't really know that the magnitude of the treatment effect is for antibacterial drugs for treatment of CAP, particularly for mild to moderate CAP.

So if there's some uncertainty--so that means that we have some uncertainty about

what is the appropriate non-inferiority margin for these studies. So just to illustrate this a little differently--you've seen some other figures, previously. In this case, in the cases of diseases that have high spontaneous resolution rates, or if there's no effective active control, there's no measurable treatment difference because the active control and placebo are about the same in effectiveness.

So non-inferiority margins would not be appropriate in this scenario.

On the other hand, if the disease has a low spontaneous resolution rate, and we have an effective active control, the treatment difference, which is here, the difference between active control and placebo is measurable. So this is the treatment effect or M1, and from there we can estimate a smaller non-inferiority margin.

So our goal, then, was to estimate the magnitude of the treatment effect of

antibacterial drugs in community-acquired pneumonia.

Usually, that's done by placebo-controlled studies, and as we've discussed before, there are no true placebo-controlled studies that we can fall back on here. So we went to the historical data on pneumonia, looking at published studies that were performed in the pre-antibiotic era, and those, shortly after introduction of antibacterial drugs.

Most have been studies of pneumococcal or lobar pneumonia, and these were synonymous at that time. Most were in hospitalized patients. Mortality was generally the endpoint that was measured.

We found some observational studies of treated patients, so treated with some, with an antibacterial, versus those that received only symptomatic therapy, or whatever was the standard of care at the time.

We also found a few controlled

trials. The treatment groups were those that received antibacterial drugs versus those that received just, again, symptomatic therapy at the time. Again, no true placebo-controlled studies, and just to reiterate, the patients weren't randomized and treatment was not blinded.

We also looked at some alternative sources of data, which might show the treatment effect between antibacterial drugs, and I'm not going to focus on this today because my focus is on the historical studies.

We did look for negative non-inferiority studies, and yesterday, daptomycin was mentioned in this context. We did not find any superiority studies. Dr. Ambrose is going to talk about studies that looked at dose response and pharmacodynamics, and this may be a promising approach, to look for some type of treatment effect.

And we looked at studies of discordant therapy. So the discordant

organisms were resistant in comparison to the antibiotic used. Whether the treatment regimen was guideline concordant or discordant, or delayed versus immediate, or broad versus narrow spectrum, empirical treatment.

So far, we've really not been able to use these types of data to satisfactorily estimate a treatment effect. Before I go into the historical data, just to quote from Sir William Osler, in 1894, who succumbed to Haemophilus influenzae pneumonia in 1919. He said that recovery followed the crisis and it brought decrease in temperature over 12 hours, accompanied by passage from a condition of extreme distress and anxiety to one of comparative comfort, occurred in a large proportion of cases. A fatal outcome was noted in 20 to 35 percent. Worse prognosis was evident in drunkards and the elderly, with fatality increasing to 50 to 65 percent in the elderly in their sixth and seventh decade.

So he's describing the natural history of pneumonia at that time before antibiotics, and we have a few observational studies which also contribute to what we know about natural history.

So just first to briefly describe the history of effective treatment for pneumococcal pneumonia. *Strep pneumoniae* was identified as the cause of pneumonia in 1881. Serum therapy, a specific anti-pneumococcal therapy, was first used with some success, starting around 1913, and was used almost until 1940.

The first antibacterial drugs were introduced into clinical practice, and sulfapyridine was the first one, around 1938-1939. And penicillin and other of the true antibiotics came into use in the 1940's. So first, let me describe the observational studies.

You've seen this data in a different way, previously. This is Tilghman

and Finland's data from 1937. So this was in a time when millions of antibacterial or serum therapy--well, for these patients, neither were used. These were untreated patients, basically.

So a couple things to note. For all cases, mortality increased with age. In bacteremic patients, mortality was higher than in nonbacteremic. The proportion of patients with bacteremia increased with age, up to a certain point in the study, about 60 to 70 years old, and I guess the other point is that mortality here, even in the youngest patients, that were nonbacteremic, was about 10 percent compared to 30 percent in those that were bacteremic.

And this is some data from Finland in 1943. He summarized some data from Boston City Hospital. Patients with pneumococcal pneumonia, treated either with no specific therapy, serum therapy, or sulfonamides. And it's a little hard to read. I apologize.

The first two groups that I mentioned, the no specific therapy and serum therapy, are historical controls here, and this was between 1938 and 1941 for the patients treated with sulfa derivatives.

So in bacteremic patients, treatment effect--or the difference between untreated patients and those treated with sulfa, was about 50 percent. Approximately the same in the oldest patients there overall, and that's probably driving this average for all ages.

On the other hand, if you look at the nonbacteremic cases, the treatment effect is much smaller, from 30 to maybe 12 percent. So a difference on the range of 15 percent here, higher for patients that were older than fifty.

This was an observational study in patients with what was described as moderate to severe pneumonia. This is by Finland's group again, Boston City Hospital. The study

was done in 1945. I'm not showing this to specifically show a treatment difference but to make a couple other points. Very few of the studies looked, actually, at severity, and using some type of severity score here, and they didn't describe in the publication, most of the patients in the study had severe disease.

So what we would describe as "severe," acutely ill or irrational, those with shock and/or heart failure. And also notice in the two treatment groups, the first was treated with penicillin alone. The second received penicillin either after failing sulfa, or in those who were intolerant to sulfa. So 16 out of 17 in the latter group, we could probably consider severe pneumonia, compared to 50 to 60 percent in the penicillin-only group.

The other point I wanted to make about this publication, this study, was that they did look at some other endpoints, other

than death. They looked at relapse complications such as empyema. They looked at bacteremia. They looked at duration of acute symptoms, duration of fever.

Again, I'm not showing this to show, necessarily, the differences between the two groups here. I just wanted to make some comments about-- we do know something about pneumonia from the historical data. And this is pneumococcal pneumonia. In patients with severe disease, mostly severe disease, mortality was 18 to 19 percent, something on the same order of what we would expect today in patients with severe pneumonia.

In those treated with penicillin, duration of acute symptoms and fever actually was resolved in less than 48 hours in 80 to 90 percent.

Here's another observational study. This is by Dowling and Lepper in 1951. They looked at case fatality rate as a function of age. In patients who received no specific

treatment--this is a solid line--serum therapy, the dashed line, sulfonamides, the open squares, and antibiotics meaning penicillin, tetracyclines.

So these two, the first two groups again are historical controls. Again, mortality or case fatality rate increased with age for all the groups.

For serum treatment, there was some benefit but mostly in the younger patients. More of a benefit with sulfa-treated patients, and even more of a benefit in those treated with penicillin and the true antibiotics.

So the treatment difference here, if we look at this, somewhere around 60 percent or thereabouts, untreated, and treated, at age seventy. A lot lower, if you're looking at younger patient. In the 20 to 29 group, the difference is only about 10 percent, and even less in those younger than that.

Here's the Austrian and Gold study,

which we've talked about some already. He looked at survival as a function of days of illness in patients treated with penicillin, and it's a little fuzzy--there was 390-some patients, and again, historical controls, serum, or untreated.

At day 21, the treatment difference is large between penicillin--and these were bacteremia-only patients with pneumococcal pneumonia. So the treatment difference was about 70 percent here.

And this slide just summarizes what I've shown about treatment effect for the observational studies. So in Finland's study, from 1943, treatment difference was about 24 percent overall. Now this includes both bacteremia and non-bacteremia patients. Much higher if you just look at bacteremia patients.

In the Dowling study, treatment difference was about 18 percent between untreated and sulfa-treated patients. The

treatment difference was even higher between untreated and penicillin-treated patients, about 25 percent.

And in the Austrian study, which looked only at bacteremic patients, and this was overall, all ages, the difference was about 63 percent.

Now I'll discuss a few of the controlled clinical trials that we found. This was a study of sero-therapy, so specific antiserum against the pneumococcus. In this case, alternate patients admitted to one hospital with lobar pneumonia, which was pneumococcal pneumonia at the time, were treated either with a specific serum to pneumococcal Types 1, 2 or 3, or the standard treatment, and here's what the standard treatment was at the time.

Fluids, pain relief with elastic adhesive plaster, restriction of opiates, no drastic catharsis, oxygen for sinus, rapid breathing, and digoxin for heart rate greater

than 120.

A couple things about this study. This is the subset of patients that had Type 1 pneumococcal pneumonia, and I'm not particularly showing it to show the treatment differences. But I wanted to show that in this study, they had some type of severity scale. They classified patients as good condition at baseline, fair, or poor, and they, again, didn't really describe it in the publication.

So in any condition--oh, the other thing is they didn't tell us how many, what the number of patients were in each subset here. But in this subset of Type 1 patients, mortality rate overall with standard therapy was 34 percent, 20 percent in serum therapy. So about a 14 percent reduction.

So if you just look at patients that were in good condition at baseline, whatever, exactly that meant, mortality was only--I say "only"--but it's interesting to

think about. Only about 13 percent, if they receive standard therapy versus 9 percent if they receive serum therapy, for a difference of 4 percent, and of course that treatment difference and mortality increased with severity.

In this study, which was also mentioned yesterday, Evans and Gaisford, in 1938, again, they studied sulfapyridine. Well, in this case they studied sulfapyridine, which was also called M&B 693. The control was nonspecific treatment, whatever the standard of care was. These are hospitalized patients with lobar pneumonia.

Treatment groups were determined by enrollment on alternate days. In this study, they excluded patients who died within 24 hours. So if you look at all patients here, there was a 100 in each group. Case fatality rate, 27 percent in those who received no specific treatment versus 8 percent of those who received the sulfapyridine. And it was

higher for those that were over age fifty.

In another small controlled study by Graham, in 1939, Graham and colleagues, again they looked at hospitalized patients with pneumococcal pneumonia. They alternated patients between sulfapyridine, and here's another name for sulfapyridine, in comparison with control. So no specific therapy.

Notice the baseline, that there was some difference in the amount of bacteremia, 34 percent in the treated group versus 20 percent in the untreated group. So the difference here was, in case fatality rate, was 23 percent for the controls and 6 percent for the sulfa group, again, a higher difference if patients were bacteremic.

And this slide summarizes the controlled studies that we found. And there's another study here I didn't mention, which I'll just mention briefly here. In this study, Agranat and colleagues in, I believe it was 1937, looked at several different

populations of patients in different locations, and they reported their results by location.

Either treated with whatever the standard of care was or sulfapyridine. And so these are the reports of two--actually subsets of subsets at one location. These were in Johannesburg, South Africa, and these were the European patients and these were the non-European patients.

The treatment difference in that study was 10 percent in this group, about 15 percent in this group, and we looked back at the Evans study treatment difference, about 19 percent and the Graham study, about 17 percent.

So to summarize, go over some of these numbers again, in the observational studies, treatment difference ranged somewhere from 19 percent with sulfonamides, 25 percent with penicillin and tetracyclines in the Dowling study. 24 percent with sulfonamides

in the Finland study, higher in the bacteremic patients.

And 63 percent in the Austrian and Gold study, but that was all bacteremic patients. In the controlled studies, treatment difference ranged from 10 to 15 percent in the Agranat study, 17 percent overall in the Graham study, 19 percent in the Evans and Gaisford study, all with sulfapyridine.

These numbers seem a little bit lower than what we saw with the observational study. Obviously, there's differences in treatment--I mean in study design, but we also think there may have been differences in severity although it's a little bit difficult to tease out, and in some of the observational studies, I show data for a more active antibiotic like penicillin, rather than the sulfapyridine.

So our point estimates for the antibacterial drug treatment effect in

pneumococcal pneumonia, in hospitalized patients, in observational studies was 19 to 25 percent, in the controlled trials, 10 to 19 percent, and higher in bacteremic patients.

There are a number of limitations, of course, to these data, using these data to estimate some type of treatment effect, and we've talked about some of these already. Differences in patient populations, such as comorbidities, immune status, pneumococcal vaccination, differences in the organism and the disease.

The old studies looked at hospitalized patients with pneumococcal pneumonia and severity was generally not well-characterized, whereas now, most CAP studies, and the outpatient study, for a number of regions, *Strep pneumoniae* is isolated less frequently but I think we would all agree that it's still the most important organism to treat.

Atypical organisms we do know are

more common, except for Legionella, in mild, community-acquired pneumonia. Clearly, there are differences in standard of care, differences in study design, differences in endpoints, differences in the study drugs. We only looked at penicillin and sulfonamides.

So I'm going to just conclude with some issues for discussion. I guess the question really is, Can we extrapolate this historical data on the treatment of pneumococcal pneumonia to estimate an antibacterial drug effect for severe CAP? Can we use it for mild CAP, or anything in between?

And then as a corollary, what is the appropriate design for CAP studies? What are the appropriate populations to study? What type of severity stratification should we use? What should be the primary endpoint? When should it be measured, and so forth?

So I look forward to our discussion on these topics. Thank you very much.

DR. GILBERT: The presentations are so great. I just wish we had more time for discussion. But we'll go on to the last one.

So we're pleased to have Paul Ambrose with us. The importance of pharmacokinetics and pharmacodynamics in predicting success or failure in community-acquired pneumonia, as well as other infections, continues to become more powerful and we've asked him to apply that knowledge to community-acquired pneumonia.

Paul.

DR. AMBROSE: Thank you. It's certainly my privilege to be presenting here this afternoon. Before I get started, I'd like to start off by thanking the organizers, and especially Dr. Douglas Webb, who's at home recovering from open heart surgery and can't be with us today. I'd also like to thank the moderators, Drs. Cox, Fleming and Gilbert, for allowing me to share with you a perspective, a PK-PD perspective on the issues that have

consumed us these last two days.

First, I have conflicts, like everyone else. I work at the Ordway Research Institute, which takes money for PK-PD research from drug companies, from the Federal Government and from philanthropic organizations.

So let me get right into it. Can PK-PD be used to predict clinical or therapeutic outcome? Or I'm sorry. Clinical or microbiological outcome in patients with community-acquired respiratory infection? And I think the answer to that, on one hand, is no. PK-PD cannot predict therapeutic response to therapy on a patient by patient basis. However, that being said, I think PK-PD can be used to identify dosing regimens, a priori, that have a high likelihood of being efficacious if--and it's a big "if"-- If we account for enough of the determinants or confounders of response in the disease state of interest.

And these determinants or confounders of response can be microbiologic, pharmacokinetic or physiologic. The problem is sometimes is we don't know what we don't know, and I think the daptomycin community-acquired pneumonia experience is particularly instructive.

Way back when Eli Lilly did a hamster MRSA pneumonia study, that demonstrated daptomycin efficacy in pulmonary infection, and based on that, in part, Cubist launched a clinical program in CAP that included two international clinical trials.

When the first of those trials completed its enrollment, it became apparent that daptomycin did not meet the criteria for non-inferiority relative to ceftriaxone, so the second trial was stopped while Cubist struggled to figure out what was going on.

And what they did is they did additional animal studies, this time using the mouse pneumonia model, using pneumococcus as

the pathogen, and this time daptomycin displayed poor activity relative to that of ceftriaxone, and ultimately, through a bunch of molecular work, they were able to demonstrate that daptomycin is bound by pulmonary surfactants, and in the presence of pulmonary surfactants, the MIC jumps a hundredfold.

This table, on the bottom, is from a paper that'll soon be published in CID, and I think is particular instructive, and I'm not going to go over all the daptomycin data and certainly "steal their thunder," but I did like this particular piece of information.

So this is a post hoc analysis of the data pooled from both studies, stratified by whether or not the patient got effective prior antibiotic therapy, effective prior antibiotic therapy was defined as drugs that have intense microbiologic activity and perhaps a long half-life, like ceftriaxone. And what you can see, if you look at the

patients that had effective prior antibiotic therapy in the daptomycin cohort, the response rate was about 90 percent, as was in the ceftriaxone cohort.

But if we look at patients that did not get effective prior antibiotic therapy drugs, like they got a dose of Bactrim, or some drug like that, 75 percent of patients had a positive response in the daptomycin cohort, relative to ceftriaxone's 88 percent.

So what might this mean? There are a whole bunch of interpretations one can make, and when you guys read the paper, you'll get to see a number of others. But what this might mean is that, remember, I said the MIC to daptomycin in the presence of pulmonary surfactants, jumped a hundredfold. AUC to MIC is the PK-PD driver of efficacy for daptomycin. That means the AUC to MIC ratio drove towards zero.

Could that 75 percent response rate be a clue, something close, something in the

neighborhood of the no treatment response rate? Keep that in mind. We'll come back to this a little bit later.

So what about clinical PK-PD analyses? Before I get into that, I think I have to acknowledge, or we have to acknowledge some of the challenges of conducting these analyses.

The first is I'll demonstrate for you, there are very few patients in these databases with exposures that are consistent with failures or suboptimal outcomes in the animal models, and further, as we talked about yesterday, the clinical trial endpoints that we've been using over the years may only provide a limited resolution of a drug's true effect.

However, today, I hope, despite these limitations, I can show you that these old data, and I'm talking about data from the 1990's, and early this decade, actually do provide us some useful information.

But first, I think we ought to take a little bit of a walk down history lane, and this absolutely has a point. Modern PK-PD research for antibacterials began with Dr. William Craig, in the 1980's, as he perfected or improved upon animal models from the 1940's and '50s, and in the late 1980's, the drug class that was "hot," the drug class that was up and coming were the fluoroquinolones, and we were very interested in Gram-negative bacteria at that time.

And Dr. Craig showed us that AUC to MIC ratio was the PK-PD driver of efficacy, and that is AUC to MIC went up, mortality, the clinical endpoint, went down, and when you had an AUC to MIC ratio of a hundred or so, all the animals got to keep walking around in their cages.

A few years later, Alan Forrest and colleagues, out of Buffalo, with ciprofloxacin in patients with lower respiratory tract infection, most of whom were in the ICU unit,

published this data in AAC, where we see the AUC to MIC ratio plotted against the probability of clinical cure or the probability of eradication. And as AUC to MIC went up, so did the probability of positive things happening for the patient.

They identified in AUC to MIC ratio a break point in this data, and that was an AUC to MIC ratio, total drug, of 125. Patients who had larger exposures tended to do better than patients with lower exposures.

The problem is this total drug, AUC to MIC ratio, was assumed by many to apply to all pathogens, all drug classes, and all patient populations. And I was in a drug company not long after this time, and what happened is during the 1990s, a lot of quinolones were coming forward, and a lot of them picked doses to achieve this kind of threshold.

It wasn't until the late '90s, and early in this century, that we began to

realize that this target of 100 to 125 did not apply to the pneumococcus, the pathogen that we're most talking about most interested in talking about today. The first hint of this came from an in vitro model by Melinda Lacy published in AAC, followed by data from Dr. Craig's laboratory, followed by human data that I was involved with.

So just to share a little bit of this information with you, this is Dr. Craig's data, six fluoroquinolones pooled, corrected for protein binding, and if you look at survival, as AUC to MIC ratio goes up, so too does the probability of the animal surviving, and at AUC to MIC ratios of 25 or 35, all the animals are still walking around in their cages.

If you take that target of AUC of 25 or 35, and apply it to the change in density, in the bacterial density in the thighs of mice, you come up with a 99 percent reduction in bacterial burden.

What about human data I alluded to? This is some data that we pooled, this is data from 121 patients pooled across a whole number of fluoroquinolones from a variety of community-acquired respiratory tract infections, and we found a CART break point. We identified an exposure break point through classification and regression tree analysis, and that was an AUC to MIC ratio of 34, very close to Dr. Craig's 25, and if you achieve this threshold, these patients, in total, had a 93 percent probability of having a positive response, while this little group down here had a 68 percent probability of having a positive response.

So why is any of this important? Well, during the 1990's, PK-PD began to grow up and began showing up in clinical trials with great regularity.

The drug class that was being developed was the fluoroquinolones, more quinolones than anything else. But most of

the quinolones were developed with exposures that far exceed these minimum exposure thresholds that we're identifying in the animal data.

Case in point. Here's garenoxacin. This is the free drug AUC to MIC ratio versus the number of patients by various AUC to MIC buckets for 96 pneumococcus in patients that had PK, in trials of CAP, AECB and sinusitis. One patient. One patient is in the bucket, where we think that inflection point may be, based on our animal priors. You see a smattering of failures, indicated here, in red, at very large exposures. With your eyeball, you certainly can't see a break point in these data, and we've tried statistically to demonstrate it, and we weren't able to find any relationship in these data published by lead author, Scott van Wart, Antimicrobial Agents in Chemotherapy.

So where do we find information? Where is there information that's helpful? I

think you have to look at failed programs and failed studies. The daptomycin I already mentioned. Faropenem. Grepafloxacin. This is where we're going to find places where we can enrich for failures. So these are data right out of the grepafloxacin package insert, and it's *Strep pneumoniae*, they'd studied grepafloxacin at two dose levels, 400 and 600, versus a comparator that's not identified in the package insert.

A 72 percent response at the 400 mg. dose level, 85 percent at the 600 mg. dose level, a hint of a dose response, a hint of an exposure response. The comparator had an 86 percent response rate. Because of these data, the FDA thought it was wise to put into the package insert--"Hey, Doc, if pneumococcus is your bug, you might want to stay away from that 400 mg. dose, the response rate isn't so great."

The question you might ask is, and I asked: Could we have predicted this, based

on Dr. Craig's data and the information I shared with you before? This is the probability of PK-PD target attainment. I just ran a 5000 patient Monte Carlo simulation and asked the simple question: What proportion of patients, given the PK, following a 400 mg. dose, would have an AUC to MIC ratio, given the MIC distribution of grepafloxacin against pneumococcus, and you can see, 57 percent of patients at the low dose achieved this critical threshold from the animal models.

If you dosed it a little bit higher, 600 mgs.--remember, grepafloxacin is nonlinear kinetics, so you increase the dose a little bit, you can get "quite a bang for your buck."

And now 95 percent of patients achieve this threshold. I bet if Otsuka had done this, had seen this data before launching their trial, they might not have allowed that dose to go forward.

In any event, that PK-PD break

point from the human data that I shared with you earlier, for grins, I tested it. How would that--what kind of response rate would that have predicted? It would have predicted there was an 80 percent response rate. The actual observed rate was 72 for the 400 mg. dose, and for the 600 mg. cohort, it predicted 88 percent probability of response versus 85. So I think the PK-PD does have some predictive value.

So why should we give a hoot? I mentioned earlier, that during the 1990's, and the early part of this decade, we've conducted more analyses in community-acquired respiratory tract disease than any other. Well, these relationships, when we identify them, the exposure response functions, they have y intercepts, don't they? And it may be reasonable to think of the y intercept as a beginning, as a place to begin to think about, Is this the no treatment effect? Is this getting somewhere in the neighborhood? Maybe

a little of an overestimation? That might not be bad. But are we in the neighborhood? And if we start looking at multiple exposure response relationships, or pooling data, so we've got more robust sample sizes, can we increase our competence in these y intercepts?

Let me show you what I mean. This is grepafloxacin in AECB. These data were published in JAC. This was part of Otsuka's program. That AUC to MIC ratio on the axis, probability of clinical cure on the y, you can clearly see is, AUC to MIC goes up. So too does the probability of response.

There's about 80 patients in this picture, by the way, and that y intercept, somewhere around 70, 72 percent; right in there. These are data that were published by Preston and Drusano, a very well-known and famous paper, involving levofloxacin for the treatment of community-acquired infection, which included a cohort of patients with skin and soft tissue infections, pulmonary

infections--that's the red line, that included patients with community-acquired pneumonia, and AECB, as well as urinary tract infections. And their model predicted, for everyone put together, stratified by infection site, look at the y intercept. Again, 72, 70 percent; right in there.

In preparation for this meeting, we pooled some data that we have access to. I mentioned, we conduct exposure response relationships regularly, and this is data from the gatifloxacin and gemifloxacin, NDA. This is pneumococci in community-acquired pneumonia. These patients were all treated orally. The vast majority on an outpatient basis. AUC to MIC ratio, probability of microbiologic response, probability of clinical response. We found--I used CART and identified a break point, came up with a similar break point to what we came up with before, 33.8 patient that had--this grouping of patients that had AUC to MIC ratio, is less

than that, had about a 67 percent probability of a positive response. These guys up here, a 93 percent probability of having a positive response.

I'm going to admit to you, right now, these are small sample sizes. We can drive a bus through the confidence intervals that are here. But these are three analyses, from three different groups, at three different times, that are non-overlapping, that are pointing us in the same direction.

And you put that in context with the daptomycin information I shared with you together. Are we getting to a place where we've got a plan to go forward and begin to do something with this?

The implications are obvious. The FDA, to date, has not found it possible to define a non-inferiority margin for active-controlled non-inferiority studies for some community-acquired infections. This is because they don't have a consistent and

reliable estimate of efficacy relative to placebo. By developing exposure response relationships, we may have a little bit of a way out of this conundrum, and if we can do it with enough patients, we really may be able to get away from doing trials that may put patients at some degree of risk in a placebo-controlled trial, or excessively low-dose ranging, or inappropriate comparators in clinical trials.

So my "call to arms," to start off with, is, you know, you guys out there industry, you guys have conducted lots of clinical trials over the last 10 and 15 years in community-acquired pneumonia, many of them collecting pharmacokinetic information.

Why don't we consider pooling? With all these quinolones, why don't we pool across complete NDA, so that we've got robust sample size, robust numbers of failures?

If we can't do that, some of these drugs really lend themselves to using

demographic models to predict pharmacokinetics in patients without drug samples, and that'll increase our sample size further but our estimation of PK certainty will go down a little bit.

And if we don't want to do that, why don't we consider using surrogates for exposure, like dose over a patient's weight, how much drug did you throw into how big a body, over MIC, and see what kind of relationships we derive there.

Maybe then we can increase our confidence in these y intercepts, or what we think may be something in the neighborhood of the no treatment effect. If we do this, and we're successful, does that mean we're done? No. I don't think it means we're done. I think the discussion yesterday really lends itself, that I believe anyways, that our endpoints are not all that great.

Looking 10 to 14 days after the end of therapy is probably not the right thing to

do, and, in fact, in the past, our forefathers, it's been pointed out in the Petersdorf study, looked at lots of things--patient sense of well-being, chest pain, appetite and cough, and note that some signs and symptoms resolve faster than the other, and unlike some of the statements made yesterday, you know, it takes some time--not everyone is feeling good after 72 hours of therapy. In fact, only about 50 percent of patients would say they were completely resolved, at least in their perception, in this database.

So if we can move away from this dichotomous endpoint, cure or failure, 10 or 14 days after therapy, and start using continuous numeric endpoints, it's more sensitive. We're going to gain power to discriminate differences between regimens that are meaningful.

This cartoon depicts the current clinical trial paradigm. We have a duration

of treatment, say, 10 days, a period of observation, we let them go a week or so, and then a test of cure window, and if we've got drug regimen A and drug regimen B, and we only look at them here, we've lost all this information, and this information is fundamental information. It's critical to the patient. They sure care. It's critical to the physician and it's critical to society.

This is not make-believe. This is not theoretical. There are examples of this. If we use these endpoints, we can evaluate the impact of drug exposure on time to event. Here's duration of treatment in the ciprofloxacin lower respiratory paper by Forrest, I mentioned earlier, versus culture positivity. Now we can clearly see that they are stratified by AUC to MIC ratio, or drug intensity. Patients with larger drug intensities tend to clear their infections faster than others.

I can show you the same data

presented at this last ICAP, where we used time-to-fever resolution in patients with typhoid fever. We used time-to-fever resolution, patients treated with tigecycline for community-acquired pneumonia and were able to stratify, at least in a univariate way, based on drug exposure, intensity.

I think if we move to these continuous numeric endpoints, we can impact the numbers of patients needed for a clinical--show meaningful differences between regimens, and this is just a table, I'll let you go through, from our gatifloxacin sinusitis data, where we looked at patients continuously, both in terms of bacterial eradication and sign and symptom resolution.

I think with this information, we can begin to define the optimal length of therapy. We certainly can get much more data from our Phase II, III clinical trials, that make a difference to researchers and treating physicians as well.

So to begin to conclude, is I think we can use PK-PD to identify regimens ahead of time, that have a high probability of being efficacious.

I also think we can use data developed from clinical studies in the last 10 or 15 years, that can allow us to get some information on what the magnitude of the treatment effect might be, and I believe if we add some new clinical trial endpoints, we can better describe drug effect and evaluate the impact of drug exposure on patient outcome, being additional information that's important to our patients and our physicians, impact the numbers of patients required for trials, and ultimately define the ultimate length of therapy.

With that, thank you very much.

DR. GILBERT: I realize this is a huge block of incredibly valuable information in a short time, and even though folks may be getting hungry, I think we have to take some

time for questions and comments.

Dr. Powers.

DR. POWERS: Mary, when you were looking through the information on the historical evidence, the one graph on the Austrian and Gold shows that it was 20 days was the timing at which you saw the large treatment effect, could you get any idea from the other ones, when they were measuring all-cause mortality?

DR. SINGER: Most of them didn't say, exactly. Most of them were look-backs at the data. It didn't say if it was just during hospitalization or not.

DR. POWERS: Okay. I want to make a point about the Austrian and Gold data too, and I'm glad you showed that graph, cause we talked about it, how many times, over the last two days.

That, as Mary knows, looking through it, is all three of those lines come from three different places; right? Austrian

and Gold's is the pneumococcal stuff, and you got the serum therapy from one place, and you got the historical control, no treatment, from some place else.

That's very useful for looking at that large treatment effect on an objective all-cause mortality endpoint. What it's not very useful for is comparing that time-to-event analysis, because you've got issues with selection bias, baseline comparability, and issues of missing data and censoring, and all that.

So we keep using that early piece of it, which is actually the most inaccurate piece of it, to say there's no treatment effect, early on, and to say we can exclude people post-randomization based on death, neither of which are appropriate.

I just want to get that because it keeps coming up.

DR. GILBERT: Thank you.

Dr. Rex.

DR. REX: John Rex from AstraZeneca. I have a question for clarification that could go possibly to Paul, or some others. It has to do with this notion that we have been defining our test of cure out at 21 days, and you've often commented, John, that we've waited too long.

But something that's occurred to me is that, yes, that's actually what you have to do to succeed. You have to get all the way out there.

But we actually permit failure at any point along the way. I can say you're a failure on day three because you've gone to the ICU. You've gotten worse.

So there is actually an implicit, I think, time to event for failure, that's built in, because you can fail early, you can't succeed until the end. And Paul, I might initially point this at you.

Do you have a sense, from your data, of when these failures occurred? Again,

I'm looking for clues, for things that we could build into design, a future Phase III based on what we know now.

Can you tell us when you think people failed? Was there any other source of time-to-failure information? Not time to success but time to failure?

DR. AMBROSE: Usually, there's one observation at some on-therapy window that's captured, usually day three to five, or something like that, and I think, to really get your answer, I think we need more observations on therapy than just at that one window.

DR. GILBERT: I'd like to take the chairman's prerogative, just for a moment, cause I see Bob wants to talk and I'm just dying to ask him a question.

So if we are stuck and we have two potential benchmarks for treatment effect, we have the historical data that Mary presented, which I found most impressive, and then

bolstered by the PK-PD data, do we have a benchmark?

Can we then create our non-inferiority margins, etcetera, etcetera, cause we have a benchmark?

DR. TEMPLE: Well, that's, to some extent, what the division has been working on for months now, I would say, and I think cautiously speaking at least, they think there probably is an effect size in the neighborhood of, I don't know, 15 to 20 percent, or something like that.

I was struck by the concentration response data as perhaps confirming the idea that there's an effect size in that neighborhood. Having said that, one always has to give the reservation. We have a long document on doing dose response in clinical trials, and it acknowledges that measuring response according to concentration is a good idea.

But it always notes that there's

some confounding. For example, maybe it's the fat people who have the lower concentrations, and maybe fat people don't do as well. Just a typical example.

That doesn't mean you should dismiss this. We have this discussion constantly with our clin-pharm people, because they love to do these modeling and simulations, and we say, yes, but you've got to keep this reservation there.

I still wouldn't dismiss those things. I think it has some further rate but you always have to keep that reservation. You know, dose response, or concentration response in a randomized trial, is unequivocal evidence of effectiveness. It's one of the kinds of trials we mention and it's perfectly good.

If it's nonrandomized, you always have to decide how worried you are about some confounding between concentration, and other factors that might affect response.

But I've found some of that--you

know, you've got to look and see if it's only the very big people who have low concentrations, or look at other factors that might predict a bad response, and if there really aren't a lot of those, I think it does add to the weight, just as you say.

DR. FLEMING: Maybe just to add to this background, it seems to me there are really, at least two issues here. Mary's presentation, as I saw it, was giving us more insight into whether there's a margin for a mortality endpoint, and Paul's presentation here seemed to be more reverse, or returning to yesterday's discussion--Can you have a margin for a clinical response type measure?

And on the latter point, it's really reinforcing what Bob is saying--there's certainly clues here but one has to be very cognizant that a concentration response is absolutely confounding the characteristics of patients that lead to various concentrations versus the treatment intention. You really

need to randomize to high, low dose, in order to be able to look at a treatment causal outcome.

And you also have to be careful if you're pooling across studies. You can look at treatment effect against a randomized control but when you're pooling across studies, now you have historical issues and then you have extrapolation issues.

So all these issues should lead to great caution. The best say, if you wanted to do this, would be to randomize dose response, randomize a low dose against a high dose and then look for the causal influence of dose.

DR. TEMPLE: But Tom, you have to bring the whole context. I mean, some of these break points are plausible because they are break points that relate the dose, the area under the curve, to the MIC. I mean, you know, things always sound plausible when you want to believe them. But that's not so crazy.

DR. FLEMING: Sure, but if I want to look at MIC and its relationship to outcome, I want to do it for a group that's randomized to a certain schedule, randomized to a controlled, lower dose, or no treatment.

DR. TEMPLE: No; no. Randomized would be great. But nobody will--that's the problem we have here--no one will let you deliberately randomize to an inadequate concentration. You have to wait for inadvertent use of an inadequate concentration and then see if you can learn something from it. I'm just saying, this is--the confounding is always a worry, you always have to worry about it, but you're allowed to sort of look at it too, and don't get overwhelmed at your ability to figure things out. But you should look anyway.

DR. AMBROSE: And I think I just--

DR. TEMPLE: And don't get overwhelmed at your ability to figure things out; but you should look anyway.

DR. AMBROSE: And I think when our datasets are big enough, we always try to look at things like patient weight as a covariate, at least, and take a look at some of these things. We don't just do a univariate analysis.

DR. FLEMING: Understood. But I come back to what makes you different from me that's accounted for by a known and recorded covariate. It's the "tip of the iceberg."

DR. GILBERT: Okay. We'll try and get in as many of these questions as possible.

Barry.

DR. EISENSTEIN: Barry Eisenstein, Cubist. A brief comment and then a general question for the panel. Going back to the data that Paul presented on the daptomycin failed trial, we see that the ceftriaxone group, in those who didn't get prior effective antibiotics, were significantly better.

But also, if you look just at the daptomycin arm, the group that got prior

effective therapy did as well with one a day or less, as the full treatment group did on the ceftriaxone. To me, this makes several points.

One of them is that there is a treatment effect, and the other, very importantly, is that you may be able to cure CAP with essentially one day of therapy. That raises the question then: How do you design a CAP study, particularly in the United States, where the vast majority of individuals that are going to enroll in a study are going to be getting prior effective therapy?

DR. AMBROSE: Can I respond to one aspect of Barry's learned comment. When the people who got effective antibiotic therapy, the largest percentage of them got ceftriaxone, a dose of ceftriaxone, and if you look at the average free drug concentration for ceftriaxonem over time, and say the MIC-50, just to get a measure of central tendency, the time above MIC following a one gram dose,

even in healthy volunteers, stretches to three to four days.

So, you know, it may be one dose but it was really effective therapy, given the half-life of that drug, of much longer.

DR. GILBERT: Next question, please.

DR. WUNDERINK: I'd just want to make a caution here. This is fine, and I believe exactly what you're saying, and I have no question, that we probably treat community-acquired pneumonia way too long. But if you don't allow us to give the one dose, you're not going to do American studies, because of what Dale Bratzler's doing, what I'm doing in my ICU. If somebody's not getting fluids and antibiotics in a short period of time in the emergency room, we're getting "dinged" as being bad doctors and bad hospitals, and we'll show up on the front of magazines.

And so you're going to kill doing clinical trials in the U.S., if you don't

allow that first dose. You know, I already have major concerns about, you know, not going to Western Europe, but going to other places around the world, and taking data there. That data may be more pertinent to the older studies on sepsis, but even the case that Dave sent me to present first, I got concerned about, cause I would have put that patient in the ICU.

You know, a 50-year-old, confusion, in my emergency department, better come to the ICU, at least for 24 hours, or 18 hours, until I get them fixed, because that's a high-risk patient, and I think that we need to be very careful--you know, I absolutely understand how it confounds the whole issue, but if you don't allow a single dose of an antibiotic, we're not going to do studies in the U.S.

DR. TEMPLE: Couldn't the single dose be the randomized two treatments?

DR. GILBERT: But then that might take more than the four to six hour cutoff.

We've got two government agencies that are coming at us from two different--

DR. TEMPLE: Why would it take--I mean, you've had to get those places that give that drug involved in the study.

DR. GILBERT: Well, yes, but there's just all the practical logistics that are involved. We have to know the patient's there, we have to get the study coordinator down there, it has to be the right person, it has to be the blinded person. I mean, it just goes on and on and on. You're using up time.

DR. TEMPLE: But it's also true that if they get a drug that actually is effective, the study is of no value in learning anything. So somehow, that has to be overcome, doesn't it?

DR. GILBERT: Yes. Well, we need to get you and the Medicare people together. All right. Anyway.

Yes?

DR. DANKNER: This is actually a

comment regarding Dr. Boucher's presentation.

At a conference I went to, a DIA conference, about a year and a half ago, I went to the CIOMS V group discussion, and they were proposing that serious adverse events that are reported in a clinical trial be unblinded to the investigators, not just the regulatory authorities, and all the pharmacovigilance people in the group thought it was a great idea, and all the clinical trial specialists rose up and got up to the microphone and said you can't do that, you will basically bias the whole trial.

And I think they backed off, but it is something to be cautious about, that there is still this concept about unblinding investigators to the regimen that the patient's receiving. One, they get the SAE reports. Their feeling is that not knowing what the patient's receiving, in the report is really not helping the investigator best manage those individuals.

DR. BOUCHER: Well, I guess it's important to distinguish between unblinding the patient after they fail, or unblinding them at the time of the SAE and keeping them in. I think both have hazards, because if I, as the investigator, find out that drug A might have caused Stevens-Johnson in my patient, I'm going to think differently from here on in about enrolling patients and treating them.

So that bias, that problem is there. If, on the other hand, it's thought that because of the potential safety implication, I need to know what he or she was getting, to take care of them now, because there was some hole that wasn't covering, you know, resistant Staph or it wasn't covering something they could have, cause I don't know the micro either, that's a different thing.

And I do think both have problems. But I guess I could see the latter much more than the former. I don't know, Tom, if you

want to comment.

DR. GILBERT: Okay. Thank you. We've got like one and a half minutes for each person at the mike.

DR. ECHOLS: Okay. First, let me compliment Helen. You did a wonderful job covering all the issues of blinding, and it really is much more than just drug and allocation and randomization.

And the key thing I want to touch on is the microbiology. We're doing placebo-controlled trials and keeping the investigator blinded, and we're doing that primarily because it would introduce huge bias if they knew if the sinus culture was positive or the sputum culture was positive.

I hadn't really thought about it in terms of CAP trials, and one of the issues is if someone has a positive blood culture, do you keep the investigator blinded from the positive blood culture results? I think these are really key issues, particularly as you get

into sicker patients, where you may have 10 percent, 20 percent, or even just patients that are hospitalized and you know what their sputum cultures are, their urinary antigen.

Is the investigator allowed to know that? And basically, traditionally, that's been the case. It's only been, I'd say in the last few years, that we've really blinded the microbiology to investigators, and I don't know how that's going to work in the hospital setting.

DR. GILBERT: Helen, if it'll affect patient outcome, don't you think the investigator needs to know that? I mean, the patient's at the bottom of this issue.

DR. BOUCHER: Absolutely, and I think we can think of several examples where you have to know. You have to know if there's Staph aureus in the blood. You're going to do a lot of things. You're going to be ordering echoes, CT scans, raising your antenna, and I would argue that you need to know if it's

Strep pneumo in the blood as well.

So I mean, I'm open to being educated, but I just can't envision getting my colleagues or my patients to "buy into" not knowing that information.

DR. GILBERT: George.

DR. TALBOT: George Talbot. Also kudos to Helen. I'd also like to sincerely thank Drs. Singer and Cox, and the division, for sharing in such detail the information base you're using to inform your decisions. It's extremely helpful to everybody here, to know what you're looking at, and I sincerely thank you for that.

In terms of Paul's presentation, which was excellent, a couple things. First of all, I think those data should put to rest some of the concerns expressed yesterday, that the uniformly high response rates we're seeing in clinical trials with quinolones are somehow a fluke or are somehow reflecting the natural history of the disease as opposed to the

efficacy of the antibiotic, and I think your data helped address that question very specifically.

The other thing I think that I'd mention is that the point was made about prior antibiotic therapy, and the question comes up also about the use of macrolides for atypical coverage and the inability to do that in the United States also is impairing the ability to do studies in American sites.

And finally, to go back to Paul's comment, I think the state of the art that you describe, of your work, should also lay to rest the thought that somehow we could design a randomized study to expose some patients knowingly to what would be an inadequate dose and some to an adequate dose.

I don't see how that could be done, ethically, with our current state of knowledge. I'd be willing to say I could be convinced otherwise, but at the moment, I wouldn't sign on to such a study.

DR. GILBERT: Thank you.

Jerry.

DR. SCHENTAG: Jerry Schentag, Buffalo. Actually, continuing on with what George just said, I'd like to thank Bob Temple for actually getting that part of the whole discussion today right.

When he said that all you've got to do is you've got to group these people--no, you expressed it well. I mean, you don't usually come to these meetings, so I'm glad you're here to talk about this.

But you've got it right. We don't need dose groups or dose randomizations in the antibiotic trials in order to differentiate the nuances of antibiotic response, particularly if that response is killing bacteria.

It's very easy, it's just as easy as it is in animal models, when you simply regroup your patients after you look at the variability in their actually achieved AUCs.

Some people have high MICs, some people have low MICs. It gives you about a thousandfold range in the data. Every time we've done that in the context of a clinical trial, it nicely separates out where they start to fail, and it's usually the lower number is around a 100, and below, and the ones above it do quite well, and the ones way above it do very fast responses, which is why it's nice to link it to those responses.

So you don't need to create a bunch of those little Phase IIs lately, that I've seen, which have a 100 patients at half the dose and a 100 patients at a gram. We don't need those. So it's fine not to. What we do need, however, is to measure the response in the course of the trial.

You have to measure the MIC, you have to measure the PK, or at least have a drug where you can trust to measure it.

And then you've got to realize that your asymptote isn't at 70 percent, where that

y axis is. Your asymptote's down there at zero in the person who doesn't respond. If you look at the cipro studies, there were people down there who didn't respond at all, and they were down near ten or zero or whatever.

If you look at the recent macrolide data we just published, they're all down there, way low too. That's in CAP and that's pneumococcus. And so we need to separate out what we know about pneumococcus from the test tube and make it work in humans, and we can do that with PK-PD. That will make all of these studies, and all the multiplicity problems you've got with all your clinical endpoints go away immediately, and I think we'll understand the system.

But it's got to be looked at from the perspective that we're trying to kill an organism here. If that organism isn't present, you know, then you can argue about whether or not you want to give the patient

something else; a placebo even. But you can't with the pneumococcus, because we know that organism very well.

DR. GILBERT: Thank you, Jerry.

DR. FLEMING: Just one clarification. You surely get the association, you're surely getting the association. The issue is are you getting information about what is the causal effect on the clinical cure endpoint? And that's really what we need to get--

DR. SCHENTAG: Yes, thanks for asking that. In cases where the clinical signs and symptoms that you're using as an endpoint, including the PROs, are linked to the organism, then you'll get an absolutely correlation.

DR. FLEMING: And they're associated.

DR. SCHENTAG: Where they're not linked to the organism, you won't.

DR. FLEMING: They're associated

but they're not absolute.

DR. SCHENTAG: Well, if the organism causes the fever, you will see the fever go away when you kill the organism.

DR. FLEMING: So what you're saying is whenever the organism's resolved, the fever resolves, whenever the organism doesn't resolve, the fever doesn't resolve.

DR. SCHENTAG: Yes. We're doing a bunch of studies now, kind a looking at that data with neural net modeling and some of the newer techniques that handle that time course. I think that will do it--

DR. GILBERT: Jerry, I've got to interrupt you just for the issue of time, and I think what Tom's getting at is, you know, the patient might have a fever due to *C diff*, or something else. But anyway, you guys can talk about it over lunch.

So we'll start lunch and *C diff*. Those go together. So we're going to restart again, exactly at 1:35, give you a quick 45

minutes. It worked yesterday. We'll make it work today.

[A luncheon recess was taken from 12:50 p.m. to 1:41 p.m.]

DR. GILBERT: In order to be fair to our speakers, as well as to ensure we end on our targeted time, I think we best get started, and we're pleased that John Bartlett is able to join us, and there's always this issue of the atypical agents, and dual therapy, or not dual therapy, etcetera, and who better to address this than Dr. Bartlett from Johns Hopkins.

John.

DR. BARTLETT: Thank you, David. I'm awfully glad to be here, and I'm also glad about my topic. So this is an issue that comes up rather repeatedly. So what are we talking about? We're talking about--of course there's a lot of atypical agents but the big three are of course Legionella with 50 species and 16 sero groups in Legionella pneumophila,

but probably the only one we can get our arms around.

Then we have Mycoplasma pneumoniae, and the old Chlamydia pneumoniae.

And the issues I wanted to raise are these. Can these agents be detected? And number two: Is there evidence that these atypical agents need to be treated, empirically, or even when we know they're there?

And the final one is: Are organism-specific antibiotic trials realistic?

Okay. So we'll start with the easy one and that is Legionella. I think what we could say is that we do have good diagnostic techniques for Legionella, and the one that's used most frequently and probably is the most realistic for routine use in most care settings is the urinary antigen test, which is awfully good for the detection of Legionella pneumophila sero group one, with a sensitivity of 75 to 85 percent. But this is the major

pathogen in the category and it's 99 percent specific. It doesn't detect the other sero groups but this is the one that's most used, easy to use, gives you an answer fast, and is widely embraced.

There's a number of other techniques. The culture is the gold standard but that takes up to seven days and therefore is not realistic. The serology is good but it takes three or four weeks, and therefore that is not very useful at the present time.

Now this is out of place. I'm sorry.

In terms of the urinary antigen test, this is the sampling of laboratory uses of these tests in Europe, and what it shows is what I just said, and that is the urinary antigen test is by far the favorite.

There are a number of other tests but that's the one that is most used, and this is probably one of the better reports because it is based on culture, and culture is really

the gold standard for this and it's probably the gold standard for the other atypicals, but nobody cultures for the other atypicals, and they do culture for this.

So this is one of these large reviews from a laboratory that's very skilled in getting Legionella and this is the comparison with the urinary antigen test and what it shows is the yield for community-acquired Legionella was 80 percent, for travel-associated or hotel-associated it was 94 percent, and for nosocomial Legionella it was much less.

The reason that these are so high is simply because of Legionella pneumophila, sero group one, as being the predominant organism in that group.

Now let's go on to Chlamydia pneumoniae, and that's a bug that's hard to get good microbiology data on. These are the data from a review recently by Maggie Hammerschlag, who has devoted most of her

career to this organism.

So the MIF that has been used through the years is probably the test that's used most frequently in various trials, and her statement is quoted from the paper, repeatedly and conclusively shown to have a poor correlation with PCR or culture, and the ten refers to ten citations that support that comment.

PCR has variable track record, but the CDC has--these are all home-grown. There is no FDA-cleared PCR technique. There was an attempt to get one a couple years ago in a big national study, but there weren't any cases, and I don't know if that means that chlamydia is oversold, or if that means that that was a bad year for chlamydia. But I've heard both.

At any rate, the FDA couldn't clear it because they said you don't have any cases, you don't have any positive cases.

What the CDC said when it reviewed 18 in-house records, that four had adequate

validity. None of these are FDA-approved. The comparison between laboratories is very poor, and actually, two of the four that had approved tests, one from Seattle and one from Hopkins, actually collaborated on a study and they decided to exchange specimens and there was almost no correlation.

So two out of the four good ones had discordance that was pretty bad. And a culture is the gold standard but it's unrealistic. So now this is from Maggie Hammerschlag's review and this is the frequency with which these various tests are done, and reported as positive, from all over the world, and you can see that the frequency of Chlamydia pneumoniae ranges all the way from 1 percent to 17 percent for adults with community-acquired pneumonia.

Now let me go on to another issue. I'm not going to address the diagnostic reliability of Mycoplasma because it's too much. I think a lot of people have the

feeling that the serologic test is adequate, and my own personal experience with it, in sending specimens to different laboratories, has not supported that. But I don't claim expertise there and the literature is all over the place.

For atypical pathogens and community-acquired pneumonia, this is an odd study, but nevertheless, it's got a couple of interesting points about atypical organisms.

This is the laboratory from the University of Louisville, atypical pathogen reference laboratory, and the report also includes not only their results but the result of the community-acquired organization database.

But the thing that's a little bit unusual about it is that this has nothing to do with this. So this is one report, this is another report, they're both in the same report, and you would think--oh, I've got everything out of order. sorry.

So these are the atypical pathogens in terms of the frequency in that report. So this is the number of cases that that reference laboratory at the University of Louisville had, and this is the percentage, and you can see in all different areas of the world, it turned out to be about 20 to 25 percent were caused by one of the three agents that I mentioned, and mycoplasma was always the top organism in the three groups, in the three categories.

In terms of coverage, it was widely variable, anywhere from 10 percent in Asia and Africa, to 90 percent in North America.

And so then the question was how often do these cause disease in other settings. This is the report by Tom File, and up to date. His claim is in outpatients, I'm sorry he's not here, but it's a good report, and it's a massive analysis of data, and it sort of fits. When you average all the cases, it's about 30 percent for atypicals and

outpatients. But for Legionella, it's exclusively, almost exclusively in the intensive care unit, it's the only one that's there, and I don't think that's going to surprise anybody.

Now in terms of the atypical antibiotic coverage, we've heard these two databases, or these meta-analyses reviewed by others at this meeting, so I don't need to belabor this.

But what they conclude is that on the basis of response to antibiotics, with or without coverage of the atypical agents, is there evidence that we need to treat them. So this is the Cochrane database and this is the review for the period, 1955 to 2005, randomized trials, adults, hospitalized with community-acquired pneumonia, and the question was atypical coverage with fluoroquinolone or a macrolide versus beta-lactam, 24 trials, 5000 patients.

And what that showed was that in

terms of mortality and clinical success, there was no statistically significant difference, but there was a trend favoring atypical coverage versus no atypical coverage.

However, in the clinical success category, when they extracted the poor quality studies, there was a dead tie. However, the exception was Legionella, and there, it favored coverage, and that was statistically significant.

This is a review, again, of meta-analysis of studies, and as Tom File pointed out, these are largely the same studies they reviewed in the Cochrane library review. So I'm not sure there is very much to add, except this now provides a relative risk, and what it shows is that there is little difference between beta-lactam versus coverage of atypicals for all of the studies. No difference, really important difference for quinolone versus macrolide.

But when you dissect out the

Legionella, there is a statistically, very statistically significant benefit to coverage when that organism turned out to be the one that was responsible.

This I've already talked about. But this is that study from Louisville, I'm sorry this is out of place, but it does show something that was kind of interesting. Atypical coverage, yes or no. Please remember that this is not tied to their laboratory, so this has nothing to do with those cases where there was or was not evidence of an atypical organism.

What they showed is that if there is atypical coverage, there was a statistically significant reduction in the time to clinical stability using a standard metric.

There was also a statistically significant decrease, by about a day in the length of stay, and these were statistically significant. But the mortality was about the

same and no statistically significant difference.

This is a noisy analysis, as you must know, in part, altered by the fact that it wasn't tied to any diagnosis of an atypical organism, and that's a point that keeps hitting us in the face.

And it hits us in the face here. So these are the Medicare data that Dale Bratzler talked about earlier, and when you look at this and then say, well, there's no difference between coverage of atypicals and no coverage, and then you look at this, then you have to say, well, there's something that's explaining this difference and part of it may be the fact that it's 13,000 patients rather than the much more modest numbers in the collected series of meta-analyses.

So what did they show in the Medicare database? Well, this was the cephalosporin, ceftriaxone or cefotaxime, given an odds ratio of one as the standard,

and this is when you added a macrolide, it reduced the mortality rate by 26 percent, and if you used a fluoroquinolone, reduced it by 36 percent. So if you look at these data, you can see, well, coverage of the atypical is probably, or possibly important.

However, that's also noisy, in this sense. We don't really know why that analysis showed the benefit of a macrolide or a fluoroquinolone. One interpretation has been that covers the atypical strains. Another explanation has been that the role of the macrolide at least, possibly the fluoroquinolone, has something to do with its anti-inflammatory activity.

So this is one of those studies, there are five of them, that show that in patients with pneumococcal pneumonia and bacteremia, a macrolide plus a beta-lactam is better than a beta-lactam alone. This has irritated us in infectious disease. We don't understand it. We don't like it. We wish

people would stop doing this.

But we're now at the point of five studies and we can't seem to shake it. So this is one of them. This is Victor Yu's study, the big study of pneumococcal bacteremia, and this is a beta-lactam alone, with beta-lactam plus a macrolide, and these are the data for the survival, and as you can see, the--well, this is mortality. The mortality was so much better--or is this survival? I'm sorry. Survival was so much better with combination therapy than here. Now these are patients with pneumococcal pneumonia and bacteremia.

So I guess what some could say is that these are dual infections. But I don't think most people in the room feel that that dual infection occurs so frequently, that it would have this kind of a dent on survival.

You'll also notice that this was a benefit in the early stage of disease, which is the first 72 hours, which is the part of

pneumococcal pneumonia and bacteremia where we've had trouble in making any impact on mortality.

Well, now let me turn to the treatment of the various agents in terms of addressing the issue of, Could we do a study of Legionella, mycoplasma or chlamydia? And I think you probably could with Legionnaire's disease. This is the first report, in 1976, which showed that those patients that got erythromycin or tetracycline did substantially better than those that got alternative therapy such as a beta-lactam.

So it was 10 or 11 percent versus 41 percent, and those of you who were practicing medicine at the time probably shared in the concern that this simply was the less seriously ill patient that we're getting tetracycline and erythromycin, and therefore this was a noisy observation.

But it of course turned out to be probably correct.

This is an observational study tat was done in Spain in the context of an outbreak, and they did have the opportunity to then retrospectively look at the treatment patients got, and do the analysis in terms of some of the outcome parameters like time to apyrexia, the length of stay and the mortality for those that received macrolides or fluoroquinolones.

And what they wound up showing was that the fluoroquinolones seemed to be superior, at least in the parameter of time to apyrexia or time to defervescence. And that was statistically significant.

However, the macrolide they used was either erythromycin, which I now think has been largely "ditched" as an adequate drug for Legionnaire's disease and they use no azithromycin which I think most people think is the preferred agent, and therefore I'm not sure that we still know whether azithromycin is as good as or better than a

fluoroquinolone.

How about mycoplasma? This is Levofloxacin versus beta-lactam and a macrolide for Mycoplasma pneumonia in children. It's a paper just published. John Bradley was the first author on it, and he's in the inner circle here, so we have to honor this study.

What he showed--this was an open label trial. The only group that we can really look at is the group that's under five, cause they're the ones that got levofloxacin versus beta-lactam. The group over five for levofloxacin versus a macrolide and ceftriaxone. It'd be interesting to find out if they thought that this was not an ethical study, to avoid that in the older kids. So the comparison is reduced to the under-five-years-old.

Mycoplasma was proven by IgM titer and the evaluation was at 10 or 17 days. And what that showed was--this is the whole study,

and then this is the subset analysis of those that had Mycoplasma.

In the entire study, they could not demonstrate a difference between coverage or no coverage of Mycoplasma, and in the group that had Mycoplasma there was no statistically significant difference.

Now Tom File, yesterday, showed one of the older reports in adults with mycoplasma, that showed big differences in terms of outcome, not mortality, but in terms of the duration of fever, the duration of fatigue, duration of cough, all the regular parameters, with tetracycline versus placebo. This would tend to refute that, at least in terms of the concept of the need to treat the atypical. In this case it would be Mycoplasma.

So what can we say about the treatment of the atypicals? Well, the need to treat, I would say there is no consensus, with the exception of Legionella.

The controlled trials are problems in terms of small sample sizes. There is no consensus on how to diagnose the other two. The meta-analyses have not been supportive. The Medicare database is supportive of the need to treat atypicals, but the reason for that is unclear because of the question of what the atypical coverage is adding to other facets of the treatment of pneumonia.

But would emphasize the fact that I think everybody in the room would say you've got to treat Legionella, if it's there, and that's not disputable.

How about atypical versus typical treatment trials? Well, I think one of the problems that we're going to encounter is in the United States and Canada, and in many parts of the world--and I'll show the slide in a minute--this might be viewed as unethical, because our guidelines say that these should be treated in both outpatients and inpatients.

So the guidelines for most, or much

of the world, certainly for Europe and North America, is that they should be treated. In terms of individual agents, well, there's also the concern regarding the frequency of these infections, which vary all the way from 1 percent to 17 percent. And the problem of knowing what you're treating. Again, the exception of course is Legionella.

For individual agents, the question is the macrolides versus fluoroquinolones, versus ketolides, and I guess that's an interesting question but I'm not sure that it's a major issue at the present time.

But the problems are the diagnosis and the sample size, and it would be awfully hard to do these except in the context of an outbreak.

Now you might be able to do it in an outbreak of Mycoplasma. That would be hard. You might be able to do it in the context of an outbreak of Legionella, and that might be easier.

Thank you.

DR. GILBERT: Thank you, John. I think the two presentations are so dissimilar. Can we take a few questions right now for John. Any questions or comments for John?

Yes, Lionel.

DR. MANDELL: John, I just wanted, in going through the data, you point out that for the hospitalized, there seems to be some evidence that treating atypicals might make a difference, and when you look at the etiology of hospitalized CAP that's not in the ICU, the atypicals make up almost 20 percent of those pathogens.

Yesterday, the point was brought up, several times, that people felt CAP, as you go from mild to moderate to severe is the same disease but a continuum. So if you apply that logic, that I think most of us agree with, then if atypicals are bad enough to get you into the hospital and they're isolated 20 percent of the time, then surely they must be

causing disease outside the hospital, and that we should be treating them.

I mean, that's more a clinical practice but it does verge on the clinical research.

DR. BARTLETT: I tend to agree with you. Where I disagree with you I think is that--I think we know that macrolides and fluoroquinolones are awfully good agents for pneumonia, in general. What I'm not sure of is that we know that that's because it's Chlamydia or Mycoplasma. I'm unsettled with the issue of how we can diagnose those agents at the present time.

DR. GILBERT: Roger.

DR. ECHOLS: Thank you, and John, thanks very much for that great review. I wanted to point out one thing and then just raise an issue.

The data by Arnold, which showed geographically, a 22 percent incidence of atypical recovery from clinical trial

database, half of that database was the same database I showed yesterday, looking at organisms across Fine classes, and the point I want to make is that fully 25 percent of the patients who had microbiologic diagnoses had an atypical plus a typical organism.

So it's not a pure--I mean, all Arnold had was the serologies for the atypical diagnoses because he did this as a reference lab.

He didn't know that some of these patients, significant, some of these patients were also infected with pneumococci, Haemophilus and other organisms. And I think that only adds to the confusion. But the real confusion to me is what role--and maybe with Legionella aside, because Legionella has real mortality--but as we've been talking today about using mortality in hospitalized patients as a way to determine the M1, there is no mortality associated with Mycoplasma and Chlamydia, which make up 90 percent of all

atypicals.

So what role does atypicals play in whether it's mild to moderate, or more serious hospitalized patients with CAP, in determining either the microbiologically evaluable population or even what the, you know, the Deltas might be or the margins might be?

DR. BARTLETT: Well, I think your comments are probably good. The concern I have are especially the ability to make a solid diagnosis of Chlamydia pneumoniae. I guess one of the reasons I've worried about that is because we have a big Mycoplasma pneumoniae laboratory, and one of the things they found is that, at least with the MIF test, which is kind of commonly used in many of these studies, 19 percent of the people that work in the hospital have acute Chlamydia infections, despite the fact that they're perfectly well.

So I'm worried about the diagnostic accuracy of the test that's commonly used, and

I think until we get a better--it may be that PCR will eventually be the winner in this. And I think it will. If I had to guess, I think it will be. I don't think the MIF test is actually going to probably survive scientific scrutiny. I expect that'll have to die.

Culture is unrealistic, because it's so arduous and so long. PCR is probably going to be the way to go with this, I would think.

DR. WUNDERINK: John, when you reviewed the Legionella data, do you think that the Legionella urinary antigen is adequate enough to exclude Legionella for clinical trials, and therefore we could use that as a way to get to monotherapy, especially in the hospitalized?

DR. BARTLETT: It may be the most realistic way to do it for many laboratories, in part, because if we need to treat fast and enroll fast, then we need a fast test, and the

urinary antigen is a fast test. The culture is probably the gold standard. Laboratories are struggling with the culture techniques, and it takes three or four days, or maybe a week, and therefore, that's not going to be very useful in the context of a clinical trial with an organism that kills people.

So I expect it'll, for study purposes, it would have to be the urinary antigen. And it's a great test. It's fast. It's very specific. Not terribly sensitive, for the reasons that you mentioned. But for study purposes, it's probably a realistic test to do.

DR. GILBERT: Thank you, John. For interest of time, we must move on, and Dave, can you help set up Bob's slides.

So while Dave is doing that, let me just mention in introduction, that over several decades, I've been fortunate to have had an opportunity with a few people, to have had ongoing spirited debates, debates from

which I've learned a huge amount. I call those few people my academic heroes, and Bob Temple is one of my heroes, and I mention this because I'm sure I'm not unique in that context. So we're delighted to have Bob here to give his perspectives on issues of non-inferiority trials.

DR. TEMPLE: Well, the trouble is almost everything worth saying about this has been said by John or Tom and Bob, so I'm not sure how much I'm going to contribute. I may go through some of these fairly fast.

So this is the non-inferiority trial story at FDA. I should start off with an anecdote. I remember when it first occurred to me that this was a problem, when I was directing the Cardiorenal Division in the late '70s, early '80s, and someone wanted to get a claim for angina by showing that nadolol was indistinguishable from propranolol in the study.

Well, we'd just been agonizing

about this for years, because there were probably 40 studies of propranolol against placebo, only a small fraction of which had been able to distinguish propranolol from placebo.

So it came to us, in a flash--if they can't tell propranolol from placebo and you don't see a difference between nadolol and propranolol, what would that mean? And then we started writing about it. It hadn't really come up much. As you'll see, other people had thought about this. Just we hadn't.

So the problem with non-inferiority trials is that they always pose inferential problems and you use them almost always--pose inferential problems, and you use them because you don't have a choice. You simply cannot leave people untreated or placebo-treated because you have a control that is necessary for their health.

Just because you need to use an active control doesn't mean the design is

going to work. There's some things, we can't tell you how to study. I'll give some examples, eventually.

So the non-inferiority study, as you've heard repeatedly, has to show that the new drug isn't inferior to the standard drug by too large an amount. What's too large? The amount is going to be called the non-inferiority margin, M, or Delta, depending on whether you feel Greek or not, and the non-inferiority margin has the two determinants we've been talking about.

First of all, the degree of inferiority cannot be greater than the whole effect of the control, because if it is, then you've lost the whole effect and that means you don't have anything.

So you have to know what the effect of the control is in the new study. But of course you're not measuring it. There's no placebo. So you have to deduce it from something else, and the notation we've been

using is to call the whole effect of the control M1, that being the largest possible non-inferiority margin you could have in a trial.

In addition, though, the inferiority must not be clinically unacceptable. And this is not a statistical judgment, it's a clinical judgment, and I have to tell you, it's always a compromise. I mean, if you have a mortality effect in a trial, why is any loss of effect acceptable.

The fact is, however, if you're too rigid, you can never have another drug. This came up with thrombolytic agents, and things like that, where we accepted a 50 percent loss of what we thought the effect was as acceptable. Very controversial. We had an advisory committee tell us it really needed to be 75 percent, but the study size to do that would have been well over 50,000.

A 50 percent reduction required a study size of 15,000, which people were

willing to do, and the question is, well, why do you need another drug if you already have one? And the answer usually is you probably need more than one drug within a class because drugs have side effects, and so on. Anyway.

So the largest clinically acceptable difference we call M2, it can never be larger than M1, that's important to remember, and people have not always remembered it.

And the critical problem in any non-inferiority trial we have referred to as assay sensitivity. Is this a trial that could have detected the difference of interest, if there were such a difference? And to do that, the active control must have had an effect in this study of at least M1. If it didn't, then showing inferiority of the test drug is less than M1, that's the non-inferiority standard, doesn't prove a thing.

If you think the non-inferiority margin of relevance is ten, and the control

drug and an effect of five in this study, and you were allowed a difference of six, you haven't shown anything. That's the problem. And since you don't measure it, it's always a problem.

You don't measure the effect of the control, so you have to assume it or deduce it, based on past experience, and if you're wrong, you'll approve a drug that doesn't really work. We don't like to do that.

This problem has long been recognized by people. One of my favorite examples. This is a citation of expert opinion. I don't give the name because the person was in the audience and I was playing with him. But this is a quote from Lou Lasagna from about 1978. He knew this all the time. He said if you can't use a placebo, you can try to compare the new drug and the standard, but that's only convincing if the new remedy is superior to standard treatment. And that's true. An active control trial

showing superiority is always fine.

If it's inferior, or indistinguishable, the results are not really interpretable, because in the absence of a placebo, you don't know if the inferior new medicine has any effectiveness at all.

An equivalent performance may simply reflect the patient population that can't distinguish between two active treatments. That's a description of a lack of assay sensitivity.

He then identified depression as a particular case in which a non-inferiority trial would not be very persuasive. And he's absolutely right. About 50 percent of all depression trials of the drugs we know and love, and we assure are effective, can't distinguish drug from placebo. That's been true for decades. Nobody quite knows why. Anyway.

We began worrying about this in a formal way, as early as 1982, when we were

rewriting our regulations on what an adequate and well-controlled study was. They were finally published in 1985. But what we said was if you are going to do an active control trial, which was one kind of acceptable control group, the regulation said, and still says, if the intent of the trial is to show similarity of the test and control drugs, the report of the study should assess the ability of the study to have detected a difference between treatments. Similarity can mean either that both drugs are effective, or neither was, and the analysis should explain why the drugs should be considered effective in the study, for example, by reference to results in previous placebo-controlled studies of the active control drug.

That's not as good an explanation as we eventually came up within ICH-E10, but considering when it was, that's not so bad. Anyway, that's the problem.

So the problem with equivalents or

non-inferiority trials has been regulatorily-recognized for more than 20 years, and of course as John and others said, the critical need to draw your inference from the past, from historical observations, gives the non-inferiority study a somewhat distressing similarity to historically-controlled studies about which we are always nervous.

There was a time when non-inferiority trials were called equivalence trials. As was said earlier, they really don't show equivalence. They really don't show non-inferiority either. But in the past--and you'll see this in publication in recent years, unfortunately--people will compare one drug with another, find no significant difference, and declare equivalence, or victory. You know, people still do that. But, in fact, as Tom said, you only show equivalence if you're better--somebody said that, and I forget who--and no significant difference can merely mean the study was too

small.

So, again, what you're looking for is that some degree of inferiority, called M, or M1, of the test drug, has to be--of the difference between the control and the test drug, the degree of inferiority, has to be smaller than some margin, and we test that by looking at the confidence interval for the difference and make sure that the lower bound is less than M1.

So as people have said, it's a not too much inferiority study, and the analytic procedures are very much like what we do in a placebo-controlled trial where we look at the difference between the drug and placebo, and demand that the lower amount of that be, of the confidence interval, be more than zero. The same practice, relatively easy to understand. So I won't do that.

Just to make the point again, which people have--everything in this depends on the validity of M. You have to be sure you know

what the active control was in the new study. That makes us inclined to choose it conservatively. You don't want to make it too high, because then if it wasn't as big as you thought in the new study, you're going to make a wrong inference, and we don't like that.

Now you've probably heard--it's important to remember that the question of assay sensitivity is not a matter of power. Power tells you what kind of difference you would have detected. The issue of assay sensitivity is how big the difference actually was between the active control and a placebo. Would there have been one? If there had been one.

So it really isn't a matter of power. The power can be infinite. But if in this trial, the active control had an effect of zero, it doesn't matter. You'll never show the difference that you needed to.

The additional problem is it's worth remembering that you only do an active

control non-inferiority trial because you're worried about leaving people untreated with the wonderful drug that you know works. Well, that means you don't really want to lose all of the effect. You want to lose a little of the effect, not too much. And there's a tension there as well.

Finally, the important thing to remember is that sloppiness obscures differences. In a trial designed to show a difference between treatments, the people carrying out the trial have infinite incentive to get it right, to be careful, to collect everything, to lose nobody, because the more sloppy they are, the less likely they'll be able to show the difference they want to show.

If you're trying not to show a difference, the incentive is reversed, and without being cynical or snotty, or anything it's just not a good situation. You don't really like to give people an incentive to not be perfect. We've been through this.

Okay. The three steps in determining assay sensitivity are, one, using historical information, assuring that the historical information, that the trials that gave you the historical information bear a reasonably close resemblance to the trial you're doing now, and that the trial is of good quality. Just briefly.

The ICH-E10 document tried to coin a phrase, I don't think it's really taken off the way we hoped it would, but this is historical evidence of sensitivity drug effects or HESDE. Catchy, huh?

That's a historically-based conclusion that an appropriately designed, sized, and conducted trial, with a specific active drug, or maybe a group of closely-related, pharmacologically similar drugs, reliably shows an effect of some defined size on a particular endpoint.

So the usual way you do that is to show that appropriately-sized, powered, well-

conducted placebo-controlled trials were regularly able to distinguish the drug from placebo. How regular, how many failures were acceptable, those are matters of art. But you don't like to see too many failures. Even one could make you nervous.

For example, I mean, what do we believe more than that aspirin reduces strokes in people who've had a prior stroke. We really all believe that. The largest trial ever done, however, of that, the AMIS trial, which is, I don't know, three times the size of anything close, went the wrong way, didn't show any benefit at all.

So if someone wants to do a non-inferiority study compared to aspirin, to show that their platelet-active drug is good for stroke, I don't think we'd buy that, even though we're quite sure it's true. So this can be a challenging thing to show.

HESDE, sensitivity drug effects, is an abstract conclusion about well-designed

trials. Assay sensitivity is the conclusion about the particular trial. So there's two more things you need to know. Oh, sorry. I have to tell you, for most symptomatic conditions, it's very hard to make the case that you know what a drug will do in a given trial.

So if you look at anxyolitics, depression, insomnia, allergic rhinitis, asthma prophylaxis, except maybe with steroids, heart failure, angina, gastroesophageal reflex disease, IBS, pain--the trial, lots of trials fail, usually not for reasons we understand. It's very hard to make a case in all of those, that a non-inferiority study would work. Even some other things, I mentioned aspirin, but, you know, we all believe post-infraction beta blockers work, but only five out of the roughly 35 trials that have been carried out actually were able to distinguish drug from placebo.

So how would we feel about a non-

inferiority trial? Not good. So it's a challenge.

There are some cases, though, that are pretty convincing, where the effect is huge and regularly seen. Heparin in deep vein thrombosis probably has a 75 percent or more effect; pretty persuasive. Treatments of certain acute leukemias, testicular cancer, huge effects, active control trials would be perfectly sensible.

The effect of a beta agonist in bronchospasm is acute and immediate, and reasonably large. We look at active control trials there, and at least for some antibiotics--I must say, I used to list pneumonia but I don't anymore. For strep throat, urinary tract infections, the effect sizes are so large, you probably would be convinced by an active control trial.

Okay. So the second major problem is to be sure that the results of the past apply to the present. Sometimes that's really

a problem. But the conclusion you're bringing forward is relevant, only if it applies to your current trial. If the new trial situation is markedly different, it just doesn't really help you anymore.

And there are many interesting examples. For example, even if you were to believe that the beta blocker data were good enough to let you use a beta blocker as an active control in a post-infraction setting, everybody who's had a heart attack now gets a lipid-lowering drug, an anti-platelet drug, or gets new procedures. How do you have any idea what a beta blocker does? Well, the answer is you don't. The situation has totally changed.

Even things we know very well, like that ACE inhibitors are good for heart failure, there's multiple studies that show that, but they don't tell you what the now routine use of beta blockers and aldosterone antagonists did. I don't know the effect size anymore. Okay. And there's lots of things

like that.

So you've got to be reasonably sure the situation hasn't changed too much, and then finally you have to be sure that the study quality isn't so poor, that you've now obscured everything. Sloppiness, in general, gives you a bias toward the null, and some of the kinds of sloppiness that we're talking about here probably don't increase the confidence interval but do provide a bias toward the null that makes you not want to believe in a trial result.

Some of these are poor compliance. Nobody takes the drug, you can't lose. If everybody crosses over, if you have a population, somehow, that improves spontaneously, didn't really need the treatment at all, a wide use of concomitant medications that works. I mean, one of the things we heard about is it turns out, if you take a good slug of an antibiotic at the beginning of the trial, nothing else may

matter. Well, that's going to make it very hard to show a difference.

And then my favorite is mixing of the treatments. If you mix up the treatments, you really can't lose. So, overall, there is a lower incentive to high quality in trials trying to show no difference between treatments, and everybody should be nervous about that incentive.

Finally, the other point is that some things we do, that we think of as conservative, which is a rigorous intent to treat, are not entirely conservative in the active control setting. They tend to give you a bias toward the null. We don't mind that in a different showing trial, because we think you should be able to overcome that, and we don't mind being conservative. It's not so good here.

So we now currently say, because not everybody agrees on this, that we like to see both intent to treat and treated cases,

analyses.

Okay. We've talked a lot about what M1 is. That's the entire effect of the drug. That's fine. We also choose M2. But it's very critical to remember that M2 can never be larger than M1, and we have not always paid attention to that. Sometimes people really only pay attention to the clinical difference they want to rule out, without worrying much about M1. Now that's okay if the effect size is huge.

Then, really, the only thing that matters is M1, and you don't have to worry, you don't need to consult your statistician, or anything. But that can be dangerous. We, in the past--and I can tell this cause it was on my watch--it was common, in cancer trials, to declare equivalence if survival inferiority of 20 percent was excluded.

And we were doing that for a while, and then one day we woke up and said, well, we don't know that the comparative drug has a 20

percent effect. So what are we doing? And the answer is we were correct, 20 percent was the difference of indifference to oncologists, they didn't really care, but it didn't provide any evidence of effectiveness because you didn't know that the treatment you were comparing it to had that effect. So we stopped doing it and now all the trials are bigger.

And there have been some comments to the same effect here, that if you rule out a 10 percent difference, that's just fine. Any difference that small really doesn't matter. And I wouldn't disagree with that. It doesn't matter. But it doesn't show effectiveness, unless you know that the active control had an effect that size in the trial.

And to a degree, this oncology experience, which is okay for me to tell, because I'm telling it on myself, was replicated in infectious disease. In several areas, notably otitis, acute exacerbations,

and sinusitis, we were accepting differences, 10 percent, 15 percent, that were probably larger than the effect of the control drug. Can't do that, and we've been reforming ourselves.

So it's very critical to rigorously define M1. Well, you've always heard this. This is a little bit like a historical control with all the problems, I don't think I'll dwell on those, and I don't think I'll dwell on that. This is an example. You don't need to hear an example.

An interesting question that always comes up is whether we want the active control to be--the estimate of effectiveness of the active control to be based on just a single drug, the one that's going to be the control in the study, or whether it's okay to pool close to the related ones, that gets you an estimate that's usually a little bit larger, so there's a very strong desire to do it. That's a judgment call. We've certainly

accepted both at various times.

When we're thinking about all these things--and this does come up in actually planning the, planning the study, there is a tendency to drive M1 toward a lower value, and the main one is you don't want to be wrong. So if you have a range of effect sizes, if historical experience says treating with an effective drug gets you a difference from placebo, from no treatment, of somewhere between 10 percent and 40 percent, we're not going to be very likely to pick the forty. We're going to be much more likely to pick the ten because that means you have no chance of improving, much less chance of improving a drug that doesn't work.

So we tend to choose values for M that are relatively low, low in terms of the point estimates we have, sometimes low in terms of the lower bound of the studies that we see. And it's also worth saying that even one failed study that was a good study is a

major problem. You don't really expect that to happen at all because that might happen in the trial, in your active control trial, and then you're going to make a mistake.

So there's a tendency for these things to be conservative, which is a challenge for anybody trying to do these trials.

Well, we've talked about this. I won't do that again.

It is worth mentioning, that if you're pretty sure that the difference that's of clinical interest is much smaller than the difference that's real, a lot of the problems you have go away. So until someone corrects me, I'm going to keep saying this. If, in urinary tract infections, the effect size is usually 60, 70 percent more than no treatment, and you're going to rule out ten, you don't have to spend a lot of time agonizing about that. That's really easy. Treating acute leukemia. There's a number of cases like

that.

Where on that spectrum various kinds of community-acquired pneumonia fit, is sort of what this is, in the end, all about. If we could be really sure that the effect size versus no treatment is something like 20 or 30 percent, and we want to rule out ten, that's going to be pretty easy. That's not so hard. if we don't know if the effect size is really maybe ten, then when you're ruling out ten, that's not so reassuring, and that's why it's going to be a problem.

And that's it.

DR. FLEMING: Questions for Bob.

John.

DR. POWERS: Bob, I had a question about--there was two papers in Statistics of Medicine of April 2006, and they actually talked about this issue that we've talked about a lot, of the per protocol versus the ITT populations. And they actually brought up something interesting.

That it's really why the data is missing, that was critical to the assessment of the per protocol versus ITT.

And one of the things that struck me, when looking at some of these trials, was that if you--you know, what we've been talking about here over the last couple days, of taking out the early failures cause they didn't get, you know, three days of drug. If you take them out, the point estimate then goes from, you know, the low 80's or high 70's up to 90, the confidence intervals start to get tighter, and it starts to sort of mislead you in the opposite direction of what you were worried about.

In other words, the per protocol ends up looking better than the ITT does.

So is this an issue of not just picking one population or the other, but actually looking closer at what goes into that population?

DR. TEMPLE: Well, that's certainly

what we're currently asking people. But before you leave that, let's say 50 percent of your population got a drug that's going to have a major effect on the outcome. You can't lose. You know? So I would say that you're probably better off, in that case, to try to drop those people out, and call them protocol violations. I mean, all this should be done prospectively, so you can reason it out.

DR. POWERS: But that goes to the issue, right, of why they were taken out; right? Cause those people are taken out for a different reason than missing data, or--

DR. TEMPLE: Yes. That's why we say you should look at both. I'm merely pointing out that most of the things that ITT analyses have in them, that per protocol don't, if you include them, give you a bias towards the null. And we don't mind that because we're conservative in the difference showing trials, but in this setting, you do mind it, a lot. But you're right. We ask

that the reasons be looked at and we like to see both analyses too.

DR. O'NEILL: I just wanted to emphasize, Bob took that sloppiness, but there's another subset of that, that it's really not sloppiness but we talked about it yesterday and today, and it's actually why Dr. Bartlett's discussion was important, and it's the sensitivity and the specificity of the classification system that you are using to enter patients.

If they don't have the disease, you dilute the signal. If you had an ascertainment of the endpoint dilution, you also dilute the treatment effect.

So both of those things bias you towards the null, and if you have that kind of situation, which is what we've been discussing yesterday and today, that's not a good place to be for non-inferiority design. Okay. So I'm just trying to reinforce--it's not sloppiness, it's just the difficulty of the

hand you've been dealt, and the value to good diagnostics, and the value to-- essentially, the follow-up on John's point is not to throw people away, after you've started the trial, you know, with two populations. But to define your entrance criteria so you do not have to do that. And that's the value to good diagnostics and not having a population that is a mixture of populations that are not likely to respond or not.

And then the other point that was made--

DR. FLEMING: Bob, before you leave that point, what you meant was not biasing toward the null, biasing you toward no difference, which is toward--

DR. O'NEILL: Biasing you towards no difference which is where you would like to go for a non-inferiority conclusion. That's the problem.

The other point that was made earlier, when in doubt, if you have an

adjudication board who's adjudicating endpoints, and they know it's a non-inferiority trial, there is no penalty to call high. In other words, if you wanted to actually say, if you wanted to say, is this a yes or a no? this an endpoint or not? there is no penalty to call it high, uniformly high, or no penalty to call it uniformly low. That too is a bad thing for a non-inferiority design because it biases you towards the conclusion you would like.

So those are probably three major real things that you have to worry about in this field, for this problem, which is sort of a subset of the sloppiness issue.

DR. FLEMING: And just on that point, that was the issue that Dennis Dixon was really also alluding to yesterday. If you have an endpoint, like a clinical cure that's a 95 percent success rate on the control, and you want to be non-inferior in the intervention, if there is in fact a tendency

to call everything clinical cure, then you're moving both of them up toward a 100 percent, which even if you are worse, you're going to miss that, and as Dennis said, goodness, if you're going to have a 10 percent margin there, you're already allowing a 200 percent relative increase. But it's even worse than that, because even if you are three times as bad, if you push everything to success, you're going to even miss that 200 percent increase.

DR. TEMPLE: I think what's critical is to remember the properties of these things. I like to say sloppiness because it's catchy. But it's really anything that is imprecise, imprecisions. It's all of those things. They interfere with showing a difference. Fatal, if you're trying to show a difference; kind of good if you're not.

DR. DANKNER: I know I shouldn't ask this question but I can't help myself. We've heard a lot about non-inferiority and we're being "beat over the head" with it, and

we're hearing good points about it. So the point that I have now is--and I heard this at the gemifloxacin advisory committee, that science changes. But since we all, I think, are coming to the agreement that there's non-inferiority when the M2 is larger than the M1 but we don't know the M1.

Why are we still having drugs out there that are approved for ABS, AOM, and ABECB, when the other companies now have a hurdle that probably most of them could never go across. And this isn't just an issue of fairness. This is an issue of public health, that we have drugs out there that are driving resistance, that no one in this room now can probably agree are actually doing anything for the patients other than a placebo effect.

So again I realize it's probably "the big elephant in the room," but it is really, I think, a major issue, and it's a big one that all the sponsors talk about.

DR. TEMPLE: Yes, it's a real good

question and I'll bounce it to Ed.

DR. COX: Thanks, Bob. You raise a point, Wayne, and this came up, in part, at the KETEK advisory committee, and, you know, I think it was actually Dr. Bradley who raised the point, and, you know, in that setting where, you know, we had a drug, we were looking back at the risks and benefits of the compound, that was certainly an opportunity to go back and, you know, look at the benefits and consider those in the context of the risk.

So, you know, for drugs, I mean, should safety issues come up, and, certainly, you know, that would be an opportunity to once again take a look at the risks and benefits, and, you know, to the more general point that you're raising about, you know, other drugs that are out there, I mean that's certainly something that we are, you know, taking into consideration and considering options.

DR. ECHOLS: This is Roger Echols. Bob, I was interested cause you mentioned

urinary tract infections and pharyngitis, are the two non-inferiority type studies you're comfortable with.

And all I want to point out is that those are the only two indications, maybe with the exception of bacteremia, because I don't know where that stands right now--but they're the only two indications that have microbiology as the endpoint, not clinical response.

So the ability to get hard data with microbiology is far easier, and even in other types of studies, to show eradication is easier than it is with clinical response.

So one of the soft points is clinical response, not--and so, you know, pharyngitis and UTIs is maybe easy from an NI point of view but it's really we're looking at different things than we do with the other indications.

DR. GILBERT: And they're actually monomicrobial, for the most part. I don't

know if Bob wants to comment.

DR. TEMPLE: No. I'm not fit to comment. I mean, in pharyngitis is partly cause we want to prevent rheumatic fever, isn't it? Isn't that why we do a microbial test? So there are different incentives. Not that you get that anymore; but you used to.

DR. GILBERT: Thank you very much, Bob. We'll be asking you to comment when we have the roundtable discussion, which is a good time to bring up the roundtable discussion, so--oh, thank you. So our plan up here is to hear from the next speaker, then take a brief comfort break, and then Tom has a few final statistical consideration remarks, very important final statistical consideration remarks.

Then we would like to go around the table. If you would let us know who has to leave early for reasons of airlines, and so forth--I've already spoken to Keith about this--we'll have you comment first. We just

want to capture everybody's thoughts.

So without further ado, if that's acceptable. So we're going to have another talk and then the break. We want to get the historical perspective on non-inferiority rate and the--I'm trying to do two things at once here--the speaker originally in the program was Eddie Power, but at the last minute, he was unable to join us, and Glenn Tillotson has kindly stepped in here.

So the perspective of industry, non-inferiority trials. Glenn is executive director of Scientific Affairs at Replidyne.

Glenn.

DR. TILLOTSON: Good afternoon, ladies and gentlemen. I'd like to thank the organizers for the late invite.

DR. GILBERT: Your voice is soft.

DR. TILLOTSON: My voice is soft. Okay. My kids don't usually say that, especially when going over the credit card bill.

I've been in contact with Eddie Power, just to try and get a flavor for what he was thinking for this presentation, and having got those thoughts, I was then met by Dr. Gilbert, the other evening, and told that whatever your thoughts were, make them short, make them quick. So that's what I'm going to try and do for the next ten minutes or so.

I, and many of the people in the cheap seats, approached this meeting with, I think, a lot of concerns. I think there's an awful lot of industry folks out there. I think we came here apprehensive, because I don't think we knew what was going to come down. We're virtually all here with our global drug development hats on. We've heard from many of our esteemed colleagues, and I think it's pretty important to note that, you know, the U.S.--it isn't versus the EU, it's actually--we are split by the Atlantic but we are trying to do global drug development programs. And from a personal point of view,

it's been pretty clear to us that the Europeans don't want any placebo-controlled trials. Nada. No. Whatever you want to put it. They're not interested, for a variety of reasons.

Another aspect is selection of a comparator agent. What may be acceptable here, in North America, is not necessarily acceptable in the EU. And it even varies within the EU. So it's getting kind of like choosing your pizza. You can't put all the different toppings on. You've got to get this figured out.

There are inconsistencies in terms of statistical evaluations, and with all due respect to esteemed colleagues here, we are seeing variations in the way different authorities view the same sets of data, which is a little odd.

And then in terms of respiratory tract infections, and I'll tell you a little bit more about that in a moment from the

industry perspective. If we want to get any RTI claims, we need to have community-acquired pneumonia as the anchor, as the foundation. Commercial aspects in today's environment. CAP--and I'll show you exactly how much the smallest opportunity represents. It's a small slice of that pie, and yet it is fundamental to our clinical programs. We don't get that right, we may as well kiss the rest goodbye.

And research investment goes way beyond the clinical trials that we are talking about. The best figures I could find were from Tufts Institute, estimating drug development in the $800 million mark. That's a lot of money, and clearly, clinical trials make up about a third of this amount.

If you take the commercial point of view, one of the things that is becoming very, very apparent is the value of the market in which an antibiotic is going to find itself in due course.

The blue line on the top is the

number of prescriptions for oral antibiotics prescribed in the United States each year. Basically, it's about one script for every American, each year, 250 million scripts a year, roughly.

But the red line represents the value, the dollar value, and the peak was around 2003, where we were looking at around $9 billion a year for the entire oral antibiotic market. It's plummeting, and as you can see, in about five years time, it's going to be worth about 30 percent of what it was. And if you're an investor, that's not a very good direction for the line going down.

And if you think about it, the total value of the antibiotic market is about $6 billion. That's less than many of the other big blockbuster, chronic drugs that are out there. It's not a great incentive, if you know what I mean.

There are similar issues with parenteral drugs, but clearly, the numbers of

this scale change significantly. But you get my message, which is what I am trying to put across. This shows you the amount of drug used for community-acquired pneumonia, about 4 percent of the oral antibiotic market. It's tiny. Clearly, that does give us the direction and the ability to go for those indications, and maybe this as well--oh, this is a wonderful euphemism--but nevertheless, this is an awful lot of investment in a small area.

So what are the challenges? There are ethical issues. I think we've already heard about comparator drugs and the variability, but how would you select comparator drugs when you've got variable resistance? I know we've heard--I think it was from Lionel Mandell earlier on in the day, how resistance differs amongst different countries. Placebo controls are not going to beat that one.

Implications on drug development.

Just looking at the mere feasibility of estimating things like--just the clinical response, what does it mean to different positions, and so forth?

There are clear implications of all of these things on drug development as a whole.

Appropriate endpoints. We've been through a lot of this, and I think it's obviously very important that we choose the right thing to look for, and then we will be guided as to how many people we need to subject to this.

But I think the patient-based assessments, there's a fair amount of history here, and Josh Metley--I was hoping Josh would have been at this meeting. Josh Metley published, over 10 years ago, a really interesting piece of work, where they actually started to look at some of the key symptoms that are noted amongst pneumonia patients, and they followed the changes amongst these large

cohorts of patients over a period of time.

Unfortunately, my question was nobody looked between day zero and day seven, and I think that is going to be the marker of where we go forward, and clearly, that has been the crux of the Lamping questionnaire, and various others, and yet it's taken us 10 years to get our head out of the sand and move forward.

This study has been shown once or twice. I show it because it's a well-conducted study, it's a European study, and they aim to look at the primary endpoint of clinical success. My sort of byword here--"plain vanilla" is the flavor of this study. Clinical success. Fine. But if you look, there is a tasty hidden streak, and this was in Fine group 4 patients. This isn't your sort of "walking wounded" on the street. These are patients in the hospital, and they're generally quite sick.

When you look at the tasty part,

there's a speed of defervescence. It's a number we can hang on people. 38.5 degrees, in fact, is that number. So we're actually getting something that's not squishy, that's not dependent upon the time of the moon and whether they've been drinking anything nice and interesting.

And you can see here, one drug was, according this analysis, better than the combination. In a group of Fine 4 predominant, you know, sick patients. Patient reported relief from symptoms. They also, in addition to that number, they look for other things that the patients were asked specifically about. Chest pain, weakness, and the sputum color isn't obviously asked by the patient but I'm sure they saw it at some stage on its way out. And clearly, they were asked in terms of their overall, how did they feel? Better. When did you feel better?

And if I remember rightly, the Petersdorf study from 1957, asked the same

question. So I'm not quite sure whether there's new science, because it's been around for about 50 years. We've only just realized that it's there in front of us, and we're starting to address some key questions.

And the impact of feeling better, and all of those good things, is that you actually want to get away from the hospital food and get home sooner, so the duration of hospitalization has been diminished as well. There are other factors that enable people to be discharged sooner as well. I acknowledge that.

The other part that I think is particularly important, Roger was just speaking about, is a couple of indications that were microbiologically absolutely definitive. Keith Klugman, earlier on, spoke about ways of detecting the pneumococcus. In order to detect the pneumococcus in this large study in severe CAP patients, over 750 patients were enrolled to achieve 77

pneumococcal cases. I think using some of the methods that Keith has suggested will help us move forward.

My only concern, and my question to Keith was, how universal, and how applicable, and how available are some of these very elegant methods, and bearing in mind where we do our clinical trial--even I now know where Podunk, USA is, because we do some of our clinical trial in places like that. How easy is it to do those sort of elegant methods in places as diverse as the different locations in the United States?

That's taking aside the fact that we do studies in Europe, South Africa, and other parts of the world.

So we have to figure in, from the industry side how do we do some of these techniques and still try and keep the overall balance of, Do we want to do these studies in order to get the approval? And I'll give you some numbers in a moment that might surprise

you. They might not.

For example, if you were to follow some of the ideas and suggestions in terms of the number of enrollees that constitute CE population, the number of patients that would comprise the MITT for the typical bugs we're looking for, in order to do these studies, and these are just ball park figures, I'm sure we can all rationalize it--but the accountants, the people that conduct an analysis called an ROI, a return on investment, this is the sort of thing they want to look at.

How much is it going to cost company X to do a study of 424 patients in CAP? 23 million.

If you want to go to the larger study, with this type of population, then it increases significantly. Even if you just have one of these studies with some other supported small pivotal study, it's going to be in excess of $70 million. That's off the bow, before you even start to think about how

you go for indications such as AECB and ABS, and so forth. Very rapidly, you start to-- again, the accountants go pale, and start to wonder what's going to be coming next.

So in the interest of timekeeping, I just thought, What have we learned? Well, as I say, I came to this meeting, and I know a few of my colleagues came with some concerns and apprehensions.

I think from my point of view, what I've been hearing is I believe the etiology of CAP, even in mild to moderate disease, I think CAP is a continuum, and that we've heard that from several speakers. It's not a different "beast" in a different piece. It's the patient that matters.

I think the new microbial diagnostics will help. But how universally available will these methods be for our studies? The course of progression of the disease is host-driven. Maybe one pneumococcus might produce a little bit more

virulence factor than the other, but by and large, it's the host that matters. We need to figure that in somehow.

Sadly, we are all aging, in some respects, and, you know, the comorbidity issues are also rising. So the incidence of CAP per se is probably going to go larger. But these return-on-investment issues still linger, and as an industry, we've got to figure out how do we move forward, and that clinical assessment alone is not enough to see any true differences. We need to be imaginative.

And I think what I've heard for the last couple of days have been some real encouraging comments.

So operational considerations, trying to summarize this, I think there are, from an operational point of view, there's a real impact of what goes on clinically in each country, and we're learning that just within, you know, the EU. Twenty-five, 26 countries

with probably 35 different ways of doing something.

And what the concerns are up to now is that by doing a clinical trial in that country, you're trying to subvert their standards of care, which is clearly not what we're trying to do.

But these things have an impact on how they perceive your studies. The etiology, I think we can do better, and with more work from Keith and the technical experts like that--wonderful. I think we need to focus on some of the subpopulations and that we need to define better from a clinical point of view.

Regulatory considerations. I've already mentioned the standard of care. Study design. These things are not globally accepted despite ICH guidelines. It's a real mish-masher there. Feasibility, I've spoken about. One of the areas that is quite nice, we've sort of "rumbled around" amongst some members of a pharma group, is that niche

indications, if we don't get certain drugs approved, we will never discover how effective azithromycin could be for GI infections, as we are learning, or atypical mycobacteria.

We'd never have really learned about ciprofloxacin and anthrax. So we have to get over some of these hurdles in order to find out whether niche indications lie in the future, and we don't have many alternatives, and unless we get some positive vibes from this type of event, I can see the audience getting thinner and thinner as time moves on.

Financial considerations. I won't "flog that dead horse." But you know what the problem is. I heard it from many of my friends out there as clinicians. They want more options to manage the increasingly challenging patients. They don't have to be better, just maybe safer, more compliance, a whole bunch of other reasons. But they need more options.

And really, antibiotics should be

judged on a totality of factors, not just efficacy.

So I think this meeting has, to me, and I think to many of my colleagues out there, has signaled some good encouraging signs. But it has to be mixed with an element of compromise. But more importantly, some pragmatism. We're banging our head on some brick walls. I came to the meeting fearing the worst. I've heard some good signs of compromise and willingness to try to move forward. But I think there's still some way to go.

How can industry contribute? I don't say industry "do it." But how can we contribute to establishing the new science, without jeopardizing the future of antibiotic research and development?

We'll take the ball. You give us the ball, we'll take it, but we're not going to take it all the way to the end zone. We need some help here. We need some blocking

and tackling. That's my knowledge of American football. Sorry.

And anyway, I think it's most important to remember, I wasn't totally optimistic, maybe it was the Shiraz that was too good last night, but I think we can move on to April 1st and 2nd with some optimism and some hope, providing we're pragmatic and we all learn to compromise. That's one industry person's perspective. Thank you.

DR. GILBERT: Any comments for Glenn before we let him escape from the podium?

Thank you very much. We'll reconvene at 3:15 and we'll hear from Dr. Fleming, and then we'll hear from everybody on the panel.

[A recess was taken from 3:03 p.m. to 3:20 p.m.]

DR. FLEMING: Why don't we reconvene, and I'm scheduled, according to the agenda here, to take one more 30 minute slot

for the last talk, and in talking to Dave and Ed, we want to maximize the amount of time that we have for the panel discussion, and so I'm going to try to instead, just maybe give ten minutes of informal comments, and that leaves us, we hope, with an hour, that we would then like to have spent going around the table, much as yesterday, where each of us takes about three minutes, hopefully keeping to three minutes, to give our specific thoughts about the scientific insights into the issues that we have listed for panel discussion.

So what I'd like to do is maybe just take informally about ten minutes to touch on an issue that is really getting at interpretation. We've spent, appropriately, a lot of time talking about issues of design and conduct of scientific and registrational trials in CAP. Issues of analysis and interpretation are also a very important part of this, and there is a multiplicity that is

inherently the case in any clinical research, and let's say we are following along Daniel Musher's insights here.

We are looking at measuring several different ways, in severe CAP, of assessing outcomes. Mortality, complications, time to defervescence, days in the ICU, hospital days, symptom questionnaires, etcetera. And we also would explore the data often to look at several different subgroups of patients, by organism, by age, by whether they had prior effective therapy, etcetera. Many other ways as well.

So suppose we've designed a trial, suppose we have a primary endpoint, and suppose that endpoint is based on the days in the hospital or complications, and suppose it gives a relatively unimpressive result.

But in looking at the data, we find a really encouraging result on survival, and particularly when we look into subgroup of older patients, we find an even more

impressive result in survival, statistically, significantly favoring the experimental therapy.

Is that reliable evidence? Are we able to conclude that because it's statistically significant, even though it's from a supportive analysis, this is something that we can rely on? And can we rely on the interpretation in terms of the estimate of the effect? And this is a classic problem, and so there are two elements to the problem, and I'd like to talk about both of them and then just give one illustration in this ten minutes.

One of them is interpreting p values. So the story I often tell is when I was an early graduate student, about 35 years ago, going to visit some friends of ours at a hospital to see their new infant in the maternity ward, in the nursery, and at that time they had all the infants together and there were 22 infants.

And I noted that twenty of them

were of one gender and two of the other. So I did what any one of us would have done. I computed a p value. Okay. And that p value was .0001. One in ten thousand, that this would have occurred by chance alone, if it was really 50/50, by gender.

So I have searched for 35 years to find someone to be the first author of a paper that I would co-author, indicating that the birth rate's no longer balanced. And nobody will do it. And the p value is valid. It is one in ten thousand. So what's wrong?

Well, in essence, what's wrong is I didn't go into the hospital with that hypothesis, where I saw something that was one in ten thousand. If I was doing an experiment, giving myself one chance, then that would be highly impressive. This was a data-driven hypothesis. In my life I would see lots of things.

I can reassure you, I don't compute p values thirty times a day, because when you

see something that's not unexpected, you let it go. And so the essence is, the main message from this is a p value is not interpretable unless you understand the sampling context from which it was derived.

And the only one I'm truly comfortable with is the prespecified primary analysis of the prespecified primary endpoint. Because if I get a two-sided .05 p value which is one-sided .025 in the right direction, I know if there's no effect, I'd see a result this good or better, one time in forty, by chance.

But if I let myself look at many things, a one time in forty is going to occur even by chance alone.

Well, I did understand the need for validation. So I went to another maternity ward, and it was eleven-eleven, and I was disappointed. But I did a meta-analysis, and it was 31-13, p value .008. So I still am looking for a co-author.

Well, the problem is if you let data generate a hypothesis, and you recognize the need to confirm it, you can't use the data that generated the hypothesis in the meta-analysis. The bias is still there.

Okay. Well, one other issue, and that is, is the estimate of effect okay? So in our severe CAP trial, where we looked at a secondary endpoint survival and found that it looked impressive in the elderly patients, a 50 percent reduction in mortality, is that unbiased? And again I'll use an example.

If any of you are fans of golf tournaments, you know that there are four rounds in a golf tournament, and there are many, many golfers. And if you look and see who is the best on Thursday--it's usually Thursday, Friday, Saturday, Sunday--who's the best? Somebody who's four under par. Does that mean that's an unbiased average of how good that golfer is? Well, then that golfer should be 16 under par at the end of the

tournament.

We'll take the next golf tournament. You look and see who's first. If they are four times that amount under par at the end of the tournament, I'll take you to dinner. And if they're not, you take me. And I'm already looking for the restaurant. Okay. That's regression to the mean bias. What does this mean? You never know the truth. You're only getting an estimate of the truth.

And that means that, in essence, any estimate is a combination of the truth and random variability about the truth. So if you have only a single analysis, you're going to get an unbiased estimate. But if you explore the data, our attention is drawn to those things that are really favorable. Our attention was drawn to the golfer who did the best, and that performance is going to be not just their true mean, those are the people who had a particular favorable outcome.

So if you then do a confirmatory

trial and expect to see something that you saw, that was exploratory, you're going to be disappointed.

So I'll give you one clinical example, and this was a recent example that's occurred in the setting of idiopathic pulmonary fibrosis, a disease for which we have no available proven therapies. Actimmune was being studied in that setting, and a placebo-controlled trial was done, and survival was an endpoint but it was listed as the seventh most important secondary endpoint out of ten. Okay. Partly because people didn't think you could have an effect on that clinically most important endpoint.

A biomarker was made the primary endpoint because that would increase our sensitivity, and that was based on FVC and AA gradient.

Well, when the study was done, it was an unimpressive result on the primary biomarker type endpoint. But when the seventh

ordered secondary endpoint survival was reviewed, the p value was point ten, a nice trend, and then when you looked in the mild to moderate patients, the p value was .004, with more than a 50 percent reduction in mortality.

And the conclusion in the press release by the sponsor is that these results are very compelling, this is a major breakthrough in a disease setting for which there is no available therapy. Well, this therapy was available in chronic granulomatous disease, and so people were then starting to use this on a very large scale in an off-label setting because of this evidence.

Well, is this reliable data or is it not? Well, eventually, it was recognized that you can't look at a subgroup analysis, post hoc, of a secondary endpoint, and view that to be a reliable result that needed to be confirmed.

So a confirmatory trial was done, more than twice the size, only in mild to

moderate patients, those patients in whom there was the anticipated major benefit, and the data monitoring committee recently recommended termination of the trial, endorsed by the sponsor, because the survival data are actually in the wrong direction.

So when these data were confirmed, it was recognized that while survival is so important, when you let the data generate the hypothesis, doing a very natural thing, which is to explore the data, you've got to be incredibly cautious to determine whether you are looking at a data-driven result, you're looking at regression to the mean bias in your estimates, and the p value, .004, needs to be interpreted in the same way that you would interpret the p value of .0001 in the maternity ward.

So what are the action items? The action items are as we design trials, it is important to have a prespecified primary analysis of a prespecified primary endpoint,

not just because statisticians are rigid, I like to think rigorous--but not just because statisticians are rigid. But because, if you're going to use statistics, if you want to interpret p values, then it is important to use them in a way that you understand the sampling context. If anyone gives you a p value, your first question should be, What was the sampling context? And then in terms of the point estimates, is that biased or unbiased.

There should be a small number of secondary endpoints, and that doesn't mean you stop there. You do in fact do exploratory analyses, but with great caution, and those, in fact, are generally best viewed as hypothesis-generating.

So if you have a single primary endpoint, John Powers has already mentioned this, the ICH guideline says it really is advised that that primary endpoint should be the one that's the most clinically relevant

and providing the most convincing evidence, that's also a valid and reliable measure, because that is, in fact, what is going to be the result that is statistically and scientifically most interpretable in the trial.

DR. GILBERT: So then to quickly follow up, my question, Tom, was if we were to do a clinical trial with an appropriate non-inferiority margin, and these were patients that were hospitalized with pneumonia, and our primary endpoint is seven day mortality, not 30 day but seven day. But our secondary endpoint is seven day mortality in those patients that we ultimately showed were infected with the pneumococcus.

DR. FLEMING: Yes, and as Bob says, is that, in fact, really your primary? Is it your intention, that you're enrolling a larger population but your intent is to really focus on efficacy in those that are truly confirmed pneumococcus?

DR. GILBERT: Well, I'm torn--yes--but playing by the rules, so to speak, I don't want to lose potential patients that are going to be enrolled in the trial, and that gets us moving back towards the mild--

DR. FLEMING: No; no. You just make that the next one after you win on the first. Then it's okay. You can test at .05.

DR. GILBERT: And that is okay? I mean, that is what Bob says, is that's a hierarchical approach. Of course what that means, then, is you have to have won on the first level, and hierarchical makes sense when you are very persuaded, that if you don't hit on the first level, you're not going to go to the second.

So a very inappropriate way to do hierarchical is to have high dose, low dose control, and say I'm spending all the alpha on high dose against control and I'm only going to go to low dose against control if high dose hits.

Okay. Well, first of all, if that's the case, there's no point in having low dose because there's no way you're going to get there unless you've already won with high dose.

But secondly, it's not what people are going to do, and this just happened. A recent study, not in this area, but another area, missed on high dose and the sponsor says I'm filing on low dose anyway.

Well, if that's, in fact, if that's the way you're intending to proceed, then hierarchical isn't the right approach. But yes, in fact, if you do intend to say the essence of this primarily analysis is really based on looking at confirmed pneumococcal, and I only want to go to the other group if I win--but the one problem I have with that analysis is if you win because you get a great result on pneumococcal, and you told me, in advance, that that's where the greater sensitivity will be, and there's likely

uncertainty as to whether you have an effect on the other group, then winning in the second level analysis in the other group, if it's entirely driven by the strength in the pneumococcal doesn't give you a label in the other group as well.

But Bob, you can comment.

DR. TEMPLE: Right. We would not give you that. You know, just one thing. It's true you, in some sense, want the most relevant endpoint like mortality to be your primary endpoint. But all too commonly, in a lot of cardiovascular settings, there aren't enough deaths, and you don't really expect that to be a suitable endpoint because there's not enough events.

So you pick a combined endpoint, death plus something plus something. Death, MI, and stroke, very popular. Our current labeling rules say that when you present those--if you win, you win. Now in that setting, it won't be uncommon to have as

secondary endpoints, usually in a sequential manner, the mortality findings.

So if you're lucky and there's enough deaths, or unlucky, and there's enough deaths, then you get that claim too and that's perfectly legitimate. But our current labeling says that in the trial section of labeling, you have to show the components of the combined endpoint.

We don't want p values on it or anything like that. But if it's a mortality plus this, plus this endpoint, we don't want any implication that you're winning on death, if, in fact, deaths are even. So that's a little tricky, that's not statistically rigorous, but we just feel it comes under the heading of full disclosure.

DR. FLEMING: The scenario you gave I think is rigorous. I mean, the scenario you gave, which is that you look at heart failure, hospitalization, free survival, as you do, and you win on that, you're going

to get a label on that, and it's very appropriate to then look to see whether or not you affected survival, and if in fact you do, because that is a hierarchical strategy. Where it's far more complicated is the IPF example that I just gave, where you put something at a lower level as your primary, because even though we all accepted that something else was the principal reason patients really benefit and want to take a therapy, that you believe there's minimal likelihood that you would show the difference you want to show.

Then when you fail on this other measure, falling back to that principal measure, while very logically, it's very logical to do so, the interpretation of this is far more cautious and far more suspect, and therein lies the essence of this dilemma that we would have.

But I endorse what you're saying, Bob. There are settings where you would use

heart failure, hospitalization-free, survival as the primary endpoint, which is a clinically-relevant endpoint, even though less profound than survival, and if you win on that, winning on survival should be labeled for that. That's the easier pathway.

DR. O'NEILL: There's another "wrinkle" to this, though. You're talking about show a different superiority trials. Well, you win and then you go down in the subgroup. Here you're going the other way around. You're talking non-inferiority trials where the win is I show no difference. And then you want to go down into a subgroup, make a "big deal" about a subgroup, and that's a tricker situation.

DR. FLEMING: And another point, just to follow up is, once you define a margin, let's say we arrive at a margin on a mortality endpoint, or something like that, that doesn't mean that that same non-inferiority margin then applies to all of your

secondary endpoints.

Technically, any endpoint will in fact have its own specific margin. So, in fact, for some endpoints--and this is what we talked about in this meeting--for an endpoint like mortality with the evidence that FDA presented today, there certainly is some considerable evidence for mortality to set up a non-inferiority margin. You may or may not choose to use that endpoint. If you choose to use another endpoint for which there aren't scientific historical data, it could be very difficult to define a margin for those other measures.

DR. TEMPLE: There is one other possibility worth mentioning. I was reminded of that by the last speaker. Let's say the primary endpoint is non-inferiority on some major outcome thing like survival, it remains possible that you could be superior. The drug that prevailed on that, and if it was non-inferior on that, you would then get to look

at time to resolution of symptoms. That could be a secondary endpoint in the same group sequential manner--or that's not group sequential; whatever it is. And hierarchical. And that is a possibility, and, you know, if it's true that some drugs work faster than others, that would be part of the claim, even if there wasn't an advantage in overall survival. Could be.

DR. GILBERT: Very good. Are there any final questions or comments from the audience?, because we're going to move forward to the panel here. Quick comments.

MR. TOSIELLO: Bob Tosiello from Replidyne. I have actually two issues that I would just like to get Dr. Fleming and the panel's comments on.

Dr. Fleming, you just talked about the problems of multiplicity, but there's also a problem that the statistical literature has called reverse multiplicity, which is the situation where either you must show success

on multiple endpoints, or the same endpoint in multiple patient populations, and that of course has an impact on the overall power of your study to be successful, if you need to show all of those things to be successful.

DR. FLEMING: That's true. If you have to show effects on two endpoints in order to win, then, in essence, your false-positive error rate is going to go down because you have to have seen both. Now they're correlated, so it's not going to go down as much as you think in most cases. But you're right. The price you pay is then you have less power to in fact see both of those effects, if they're real.

DR. TEMPLE: There have been suggestions in the past, that we ought to adjust a little, maybe .07, but we haven't done that. But I think the fact is if, you know, you had multiple endpoints like that, and they were all pretty close and one was .052, we might be able to survive that.

DR. FLEMING: Indeed. But of course part of the reason that it's been a difficult thing to formulate what the adjustment would be is that they're correlated, and part of the reason too is that if you have two such measures, in many cases it's because, from a scientific and regulatory perspective, they're not equally clinically meaningful.

One of them might be a great biomarker. The other one might be a direct tangible measure of clinical benefit, and therefore it becomes important as to what the relative clinical importance of those two would be.

DR. FLEMING: We do that in migraines. In migraines, you have to win on pain. You also have to win on phonophobia, and stuff like that.

So the practical thing we do is, if one of the studies doesn't show all of those things, that's okay. You know, we'll live

with that.

MR. TOSIELLO: But from a sponsor's perspective of trying to plan a study, and if you're writing guidelines on these things--

DR. TEMPLE: Yes; that's hard.

MR. TOSIELLO: --it would be much easier for us to know, that if you're going to look at these other endpoints, let's say at a alpha level of .10 or .07, as you say, I can adjust the sample size to have 90 percent power to do that, rather than softer statements that say we want things to go in the same direction.

MR. TOSIELLO: Yes. I think we'd rather you adjust your sample size, so you can win on both.

DR. GILBERT: We need to move along, folks.

MR. TOSIELLO: Okay. One other comment, please. Dr. Temple made the point that it's hard to establish an M1 as any larger than the lower bound of the 95 percent

confidence interval of the treatment difference, and I would just offer that the likelihood of the true treatment benefit being the lower bound is the same as it being the upper bound, and neither of them is very likely and it's probably somewhere in between.

So it seems like a very conservative approach.

DR. FLEMING: It may seem conservative. I would argue it may seem conservative. The first issue, though, is that part of that adjustment is reflecting the fact that it's--what I was saying a little bit--we never know the truth. We're only getting an estimate of the truth.

And so there is going to be an inherent penalty, so to speak, that would be used in any statistical method that accounts for the variability in that historical estimate.

But then using the lower limit of the confidence interval is at least, in my

view, the justification for that is the uncertainty about the validity of the constancy assumption, and while that could be an over-adjustment in settings where the constancy assumption is truly valid, it could be an under-adjustment in cases where it's truly not valid.

DR. GILBERT: George.

DR. TEMPLE: Can I just add one thing. We are, however, sensitive to the fact that it could be highly over-conservative, and we're in the process of writing--we're eventually going to write guidance on all this stuff, and I think one thing that we're going to have to come to grips with is when you might be a little less conservative. That is, when would you not take the lower bound, and the kinds of things that--I've put this on slides from time to time, so it's nothing novel.

The sort of things that might convince you is, you know, a total

understanding of the mechanism, a drug that's the same as multiple other drugs in a class. Maybe you don't have to be as conservative in those cases. But sometime in the fall, we're going to have a guidance out.

DR. POWERS: But the more evidence you have, the narrower those confidence intervals get. So the lower bound moves too. So it's not just taking the lower bound. It's where the lower bound is.

DR. GILBERT: George, very quickly.

DR. TEMPLE: That's true, but you have to decide what interval to use and how conservative to be.

DR. TALBOT: I have two important comments, one on endpoints, one on population. I'll start with population. Looking ahead to the panel discussion, bullet one, the question is, What constitutes severe CAP?, etcetera. If we go back to Scenario 2, we see that's defined as CAP pneumonia requiring hospitalization but not requiring ICU care.

That scenario being defined as severe, is different from the way clinicians are defining it, at least as I heard Lionel and others.

So I would ask you in your deliberations to consider if there's a way to make the population definition for clinical research and regulatory purposes consistent with the current approach used in clinical care. I think that would have benefits. It would have benefits in clarity for enrollment into clinical studies, and it would also have benefits in terms of the labeling and the way the drug is actually going to be used.

And a good example of that, right now, is that a lot of clinicians, here and elsewhere, don't understand what complicated skin is. They think that means severe. So my proposal is that severe be restricted for clinical study purposes and regulatory, severe be restricted to that type of patient who is perhaps PORT 5, or at least requires ICU care.

Mild, we heard people are going to find generally is PORT 1, but there are some exceptions, and then moderate can be somewhere define in between. But I'd caution you that what clinicians mean by severe CAP is not what you've defined in Scenario 2, and this is an opportunity to resolve confusion over that, and I think there'd be some benefits.

And Scenario 1 might be better termed mild, for example, and there's a Scenario 3, perhaps--well, you get my point.

The second thing is endpoints, and I get a sense that there's some discussion here, or focus on mortality, and I understand all the reasons for that, and I think that you're going to debate and discuss the merits of that.

I'd just point out to you the "elephant in the room," that there are a number of sponsors out there who have completed and filed trials with a different endpoint, I presume, the old ones, and there

are sponsors with ongoing trials, or just-completed trials, where mortality is not an endpoint, and so the question comes up, well, what happens with all of those.

DR. GILBERT: Thank you, and I'm sure the panelists will consider those remarks as we go around the room.

John.

DR. BRADLEY: A very quick question of Drs. Temple and Cox. As we talk about scientific outcomes of studies, and time to resolution of symptoms, we saw in one study that the fever resolves in 3.2 days versus 3.7 days, and we've also been talking about clinically meaningful benefits of treatment, especially as it pertains to moderate as opposed to severe disease.

How does the Agency define clinically meaningful? because I would bet that a statistically significant benefit would not always be considered clinically meaningful.

And so what we need to know is what is clinically meaningful in order to power the studies, to capture those endpoints. DR. TEMPLE: Ed's got to tell you what's meaningful in any microbial disease, but I just want to make one comment, and that is sometimes it helps to look not only at the mean difference but at the distribution of differences as well, and you can sometimes get a better feel for what a mean difference means when you do that. But as to how many hours is meaningful--

DR. COX: Yes. And I don't know that I can give you a set number of hours today. But, you know, the general impression that what you're doing is something that's of value to patients. I mean, it's something that patients look at as being an important improvement, and I realize that's not a precise answer, but, you know, I think it's sort of a general concept of how you, you know, how to look at this issue of, you know,

what's important.

DR. BRADLEY: Well, I know it's a general concept but if drugs are going to be approved based on meaningful, then I would suspect that for each disease entity and each indication, meaningful could be defined.

DR. COX: Yes, and, you know, typically the way this would come up would be, you know, in essence, either, you know, in writing a guidance document or in looking at a protocol. If somebody would come in with a proposal and propose an endpoint, we'd get a chance to look at it, comment, and similarly, you know, in a guidance document, putting together you know, the types of endpoints that we would be looking at, and that's, you know, probably the way to tackle that problem.

DR. GILBERT: Matt, the co-chairs are telling me we are literally out of time. Can you do it in ten seconds?

DR. WIKLER: Absolutely. This is very fast. I hear some talk about maybe just

considering patients who have pneumococcus, for example, as one alternative, and just to put that in perspective, we saw some numbers about how much it costs to develop drugs, and the numbers up there were based upon studies looking at all-comers basically.

So I think one needs to realize if the criteria now becomes studying or evaluating only patients who have pneumococcus, the costs of those studies go up somewhere between four and fivefold to conduct those studies. So I just wanted to make that point as you're considering the future.

DR. GILBERT: Thank you. That's why I asked the subset question, actually.

A few quick announcements, all of which I think are very positive. The Infectious Disease Society of America was strongly motivated to move forward with this workshop, because we felt that the workshop and the dialogue that you've heard over the last two days would be a giant step forward

towards removing many of the uncertainties that are involved in the development of drugs for community-acquired pneumonia, and I'm hopeful, and very positive, that I think that we've made some giant strides along those--to meet that objective.

You've already heard that there will be an Anti-Infective Advisory Committee meeting on April 1st and 2nd, and I'm now cleared, I guess, to tell you that the IDSA has been invited as a guest speaker to that meeting.

Furthermore, we're going to present a position paper in advance, that'll be distributed to all of the panel members that are on the advisory committee, and then Brad and I will try to distill down--I'm not quite sure how we're going to do it--these two days of very "meaty content" to present at the time of the committee meeting, and answer questions, and so forth.

Now with that, we have one last

chore, and that's to make sure we get all these gems of knowledge out of all of you.

So we're going to go around the table, and we put, mainly to stimulate the panel, the three questions that are in the program, what constitutes severe community-acquired pneumonia? what superiority and non-inferiority designs, what is the appropriate primary analysis populations for a trial of severe community-acquired pneumonia and is it influenced by the antimicrobial spectrum of the test drug?

Now those are the big issues, but you're to feel free in your three minutes, and we are going to time you, to comment on anything else that you feel is absolutely of critical import.

I'm going to start with Keith because I know he has to leave to catch an airplane, and then if anybody else has that problem, you'll let us know.

Keith.

DR. KLUGMAN: Thank you. I've really enjoyed this meeting, and a lot of additional insights which have been really valuable. I thought that the historical analysis really does give us a footing to define M1, as it's being defined in this meeting. I think that for the discussion today, we really are talking about a non-inferiority endpoint, and it's been made quite clear to me, and everybody here, the enormous problems we face with the types of trials that have been done up to now, and the kind of perverse incentives to introduce patients that are sloppy in their recruitment, sloppy in their follow-up, and how these things perversely all contribute to making that non-inferiority easier.

So I think that precision is perhaps the order of the day, and we'll talk to that in a minute.

In terms of the outcomes, the M1 that was defined was based on a difference in

mortality, but I'm afraid I think it's naive for us to suddenly start talking about mortality as an endpoint in these trials, when the analysis of all the past drugs that have been licensed for this indication gives an overall mortality in those studies of between 2 and 4 percent.

So I think without inordinate change in the number of individuals recruited, this endpoint is going to have to be a mortality plus endpoint, and I've got a sense that there may be some sympathy to that, so there would be a clinical endpoint.

I also have felt very unhappy about our current endpoints, which are physician-driven, success or failure, one has no idea on what basis they have made that decision, so I think some kind of a score-based endpoint, together with mortality makes sense, and even perhaps the patient-reported outcomes as part of that.

I'm also encouraged by at least a

consideration of moving the endpoint of these trials forward, because I think a lot of what we've heard is that if patients are going to get better from severe pneumonia, you're going to see outcomes within at least, or at most, perhaps a ten day horizon. And it may help industry not to have such long follow-ups, but if the--I think that we'll learn a lot more if we look in much more detail at kind of ten days, not only clinical but also I wouldn't throw out some of the microbiological endpoints that could be looked at, be it time to clearance of sputum, be it some of the new things, if we have quantitative analysis, one may be able to see differences in these, by day, over the first ten days.

Then the final contribution that I want to make is that it will contribute to precision, if the population under study have the disease that we're trying to treat.

As I said earlier, the drugs are discovered and go through phase 1 and 2 on the

basis of their ability to kill the organism. I fully appreciate that if you have a drug that kills the organisms and kills patients, that's not a good drug. But we really need to try and recruit the population that have the disease that the drug is designed to deal with and so my hope would be that we could come up with pneumonia trials where we're able to define at least the etiology in a large fraction of the population, and this will add precision to these studies, and hopefully allow us, then, to have more confidence in these non-inferiority designs, and perhaps even some superiority. Because one of the big frustrations--this is really my last point I want to make--is that we've had endless numbers of comparison studies to drugs where we know there's lots of resistance but we've never been able to show it.

So perhaps if we enrich the group of pneumococcal disease, even if we don't have susceptibility, the susceptibilities of those

that we don't know ought to be the same as what we know of those that we do culture, and I think in that scenario, we may find that some of the less-active drugs, one can actually begin to see some superiority studies.

DR. GILBERT: Thank you very much. Is there anybody else that has to leave early? We'll be done in 45 minutes, probably. Okay.

Yesterday, we started there. Today, we'll give the NIH a head start. We'll start over with Dennis.

DR. DIXON: Thank you very much. Once again, it's been a pleasure to participate in your workshop. I have just a couple of points I'd like to make. One is that, really, in follow-up to the comments just made, I think it's fine to talk about having more precision by refining the population to be studied. But I think, ultimately, it's important to study the

population that'll be treated. And so that should be kept in mind in deciding on what sorts of people should enter the trial.

I wanted to make one comment about the subset question that they've raised earlier. Of course if the idea is that you'd like to see whether you could qualify for licensure, both from the pneumococcal subset, and on everything else, the complement of that, then, in some circumstances you would just do two trials.

And if you did two trials, then this whole question about whether you have to take in--you know--make multiplicity adjustments, almost entirely goes away. Not completely but almost.

And so then the question is if you really do have those two objectives, then should the fact that you addressed both those objectives in the same trial require any kind of adjustment, when it wouldn't, if you were doing separate trials.

And, you know, we're not talking about a case where you have two different experimental treatments and you're comparing them to the same control group. That introduces a need for an adjustment that's on an entirely different basis. So I think that's worth thinking about too.

Well, and the only other point I would raise is, I think the only aspect of non-inferiority trials that hadn't been talked about at all, is the question of what would be the implications of having very close monitoring to protect against a negative outcome. In other words, that the experimental drug was actually making things worse.

That there's intense monitoring to catch that as soon as possible. Would that have any implications for the discussion about margins and how big the margins can be, and so on, because that would, in effect, represent an extra protection against, you know, too

often reaching the conclusion that the candidate was just as good as what was already available.

And I can't think of anybody better to put that question to than Tom. Thanks, again.

DR. GILBERT: Thank you. I think some of those issues of safety or adverse effects, including lack of efficacy, were addressed by Dr. Talbot yesterday not quite to the depth that you describe--or same way that you describe.

John.

DR. POWERS: So I think that the data that Mary presented show that you can justify a margin in severe community-acquired pneumonia but it hinges on three things. One is the thing George Talbot brought up. You have to define what severe is, and what that really means is defining the population. It also means defining the endpoint and also means defining the timing of that endpoint.

And let me just sort of address those things. When you look at the historical data that Mary showed, it seemed that it was people who were older, had comorbidities, and those kinds of things, that had the highest mortalities.

Now my reading of the Fine paper on the pneumonia severity index is that's exactly the things that they put into there, to look at an all-cause mortality endpoint. So it seems like that would be a useful thing, to try to select people.

Can you pick a timeframe in which to do this? That's tougher, when you look back through this information. The Austrian and Gold paper that has those three curves, and again those three lines come from three different places, but it seems to plateau out at about day fourteen, where the difference is separated.

Earlier, for an all-cause mortality endpoint, you know, there's just no evidence,

to say one way or the other, whether an earlier endpoint than that.

Secondly, lastly, the margin would depend upon what population you study. I can easily see that what's going to happen is we'll start out talking about a population that's older and severely ill, and what'll start coming in is people under forty, who are less severely ill, with a mortality of 2 percent, and then we won't know where we are anymore.

So it's going to be key getting into these trials the people who are severely ill, and you might want to set some bar for what all-cause mortality is in that setting so you've got that population.

The last thing is, I think somebody brought it up already--just because we use all-cause mortality and non-inferiority on that as the primary endpoint, still doesn't mean that we can't look at secondary superiority endpoints on things like time to

resolution of symptoms, because even if we have the endpoint at day 14, it doesn't mean we have to treat people for 14 days.

And one more rational way to do this is to treat people not at a fixed duration, but treat them until they get better and stop, and then we'll actually be able to label an average duration of treatment with a range and we'll give people what they need, cause the old joke and idea is how long do you treat somebody. Well, long enough but not too much. Right. So actually, we can do that in the clinical trial.

The analysis population we talked about, ITT and per protocol, seems like you got to look at both, but we've got to fix the people with being excluded from the per protocol. Excluding people inappropriately just doesn't make a whole lot of sense, and we've got to fix the priority therapy problem, cause it seems like at least one dose is having an effect on the outcome.

DR. GILBERT: John, I can't resist. You're turning into a doctor. I mean, that's how we treat nonbacteremic pneumonia. We treat till they're afebrile three days and we quit.

DR. POWERS: Well, you know, we are trying to make these clinical trials more closely related to what, you know, give you some answers--

DR. GILBERT: It's taking so long to get there.

John.

DR. POWERS: I've been a doctor for a long time, Dave, by the way.

DR. BRADLEY: A lot of these points have been made, so I am certainly not going to dwell on them. Tighter enrollment criteria I think just makes so much good sense to me. Non-inferiority trial designs, and as John mentioned, you know, the margins, how to determine the margins, and yesterday, we briefly talked about looking at previous

studies in which there were failures from adequate dose exposures, and Paul showed some beautiful data on modeling which I think really makes the point and I'm sure there are other studies out there in which an adequate drug exposure will help us determine what the placebo effect really would be.

And once we get these data, which are admittedly retrospective, Tom, and not the best data, as we move forward we can hopefully collect data prospectively on a much-better-defined population to determine what the treatment benefit or treatment effect would be. Patients treated with resistant organisms, where there might not be an effect of the drug, although many of these patients will get better spontaneously.

And the comment was made earlier about obese patients and lack of exposure, and lack of an adequate exposure and Dr. George Drusano presented data at the IDSA meetings on intraabdominal infections, where the failures

happened to be in obese patients. So there are data presented on exactly that concept, and those failures could be extrapolated as placebo-treated.

The populations to study. Certainly serious CAP and using mortality as an endpoint makes sense, but once you move down to moderate, where the mortality rate is so low, time to resolution of symptoms and what is a meaningful benefit, and that's where we get back to my question of Ed, which is to be defined. So I'll stop there.

DR. AMBROSE: I guess I'll build on some of the comments that John Bradley made, or Dr. Bradley made. I think there already is a tremendous database in the last 10 or 12 years with some of the clinical trials that have been done, and if we can get that data into one place, and analyze it maybe with a new way, with a view towards exposure response, we might be able to answer some of the questions that plagued us, especially in

the discussions yesterday, with inferiority margins, and so forth.

And I encourage interested parties, pharmaceutical companies, and the FDA, to help work through these problems. Glenn Tillotson, Dr. Tillotson asked how pharma can help. Provide the data. You know, it's a lot of money and it's a lot of work to get archived data. But provide the data, even if it is somewhat embarrassing in a program or a drug study, or maybe the drug didn't do so well.

And maybe this is a place IDSA can help and be the fair arbiter to get this, and to maybe help come out with the analysis plan, and how this would look like in conjunction with our statistical colleagues, our clinical folks, and our pharmacometric folks, not just those who do exposure response in humans, but those also who understand the meaning and the limitations of the animal data, which is the basis for so much of what we do.

With regard to future studies, get

PK in as many patients as you can, and also use these new endpoints that we're talking about, or these old/new endpoints, whatever they are, time-to-event analysis and these things can be quite powerful in making our future studies better. That's it.

DR. GILBERT: Lionel.

DR. MANDELL: Again, I just want to thank you for allowing me to take part in this. I found it very interesting and very enjoyable. The only downside I would say is when I left work Wednesday, if somebody had asked me how to do a trial in pneumonia, I think I could have given them an answer. Now I'm not so sure. But at least I know now what I don't know, which is a step forward.

There are a few items I want to comment on. One is the endpoints. I do think that for the serious, although not necessarily the severe ICU cases, mortality is important, and it's pretty clear, we should be looking at it earlier on, say around seven to ten days

rather than thirty. I also think that if mortality's looked at, all-cause mortality is important.

I think other endpoints, as other people have commented, are important, and as Dr. Cox said, it's what important to the patient, and many of us took the Hippocratic oath before we took an oath to become investigators.

So it is mixing clinical issues and research issues, but that's the real world. So I think other endpoints are important and I think the PROs are clearly, I think, the way to go with a lot of these. In terms of etiology and diagnosis, think some of these new methods are very exciting, but realistically, I think it's going to be quite a while before they're in many of these centers and can be used routinely.

I think also in terms of benchmarks for effect, the historical data was great, and I think that's helping us, and the PK-PD data

as well.

Another area that's come up, and we've wrestled with this with the guidelines, is the treatment for atypicals, and John did a great review on this, but I certainly think we need to treat them in the hospital. It's the outpatient group that there may be a more legitimate question about.

I would also like to make a plea for--I've always been very jealous of the cardiologists and their large-scale trials, and in ID we are always taking, you know, 30 trials with small numbers and doing a meta-analysis, and somebody once said meta-analysis is to analysis what metaphysics is to physics.

And there's some problems sometimes with meta-analysis. So I think that if industry is going to sponsor these trials, we should go for the large-scale trials rather than the multiple small ones.

And Rich asked me to say one thing for him, and he felt it was a critical issue.

That a single dose of empirical antibiotics be allowed for clinical trials in order to be viable in the U.S., and that would be true in Canada as well. Thank you.

DR. GILBERT: Thank you.

Mary.

DR. SINGER: I think it's been a great discussion over the past couple days. It's been very helpful, and I wanted to thank everybody for coming and everyone for speaking. I think there is still some uncertainty about treatment effect but I think we're at a place where we have to make some assumptions, and I think the preponderance of the data shows that there is a sizeable treatment effect, at least in severe pneumonia, or in bacteremia patients, or in older patients.

So that we can probably--and that effect is probably large enough, that we can estimate a non-inferiority margin that's clinically acceptable.

It means we would have to study patients in the same category, patients with severe pneumonia or patients that are older or bacteremic. What does it mean for mild to moderate pneumonia, or just mild pneumonia? Can we use the same margin? I think that's still a question we haven't answered.

I mean, we all think of pneumonia as a continuum of disease. But does that mean that the treatment effect is the same in both? I don't think so. So we might need to think about some other approaches, and any of your thoughts on that would be very helpful as far as how to determine the margin for patients with mild pneumonia.

I think the endpoint should be a combined endpoint, including mortality. As far as defining severity, I think that's something, I agree it's something we need to define more clearly. I don't think we have the answer today as to what the best way to do that is as far as clinical trials, but I think

further discussion on that will help. That's all. Thank you.

DR. SPELLBERG: Well, I think the historical data were very impressive, and I actually, if you apply the idea of using a PORT score equivalent to look at those patients--I mean, you can't actually apply a PORT score but you can sort of theoretically look at them in that context--you would think that, you know, the 10-to-20-year-olds and the 20-to-30-year-olds would not have had nearly as high a PORT score as obviously the people at the other end of the spectrum.

And the lowest mortality rate that I saw in the pre-antibiotic data that Mary presented were on the order of 10 to 15 percent, and I do not believe that those patients are the equivalent of what we would today consider severe.

So I think that there is a signal in the historical data, that even for moderate, I will call them moderate, not mild

per se, but there is going to be a substantial treatment effect, if you're talking about 10 percent mortality pre-antibiotic and less than 1 percent mortality in the antibiotic era. And that's completely--all that historical data is concordant with the in vitro data, the preclinical data, our understanding of the mechanism of action.

So there's tremendous concordance of the data, even though the quality of each individual piece of data is obviously not quite ideal.

So I really agree that we're talking about non-inferiority, and I agree that we're talking about for the more severe patients, mortality with other components and a composite endpoint that should be time-to-event based.

For the moderates, I still think it might be possible, using some of the information that Roger Echols described, to come up with a non-inferiority margin,

probably more in the time-to-event type endpoint rather than based on mortality.

DR. GILBERT: I won't take a lot of time to reiterate what everybody has said. My major anxiety, on a personal level, of coming into these two days, was that we weren't going to get to a benchmark, and I think thanks to Mary's efforts and Dr. Ambrose's efforts, that we have what I think the Agency wanted, which was a scientific basis, as best we could unearth it from the available information for a benchmark, and I think we're there.

I certainly agree with the comments that precision and definition of mild, moderate and severe is needed, and obviously is part and parcel of the entire enterprise. We certainly need these partnerships. I know I'm harping on the same thing as yesterday, but between the companies that are developing these rapid diagnostics, hopefully some day point-of-care test, and the drug company sponsors, I mean, the potential for synergy

there is just incredible.

And then lastly, what I'll just euphemistically call creativity, creativity in the way we characterize the time-to-event endpoints. I think we're taking baby steps and we're going to learn a lot as people explore different ways to do that.

DR. FLEMING: Thank you, sir. Well, I started, as I do many times in design of trials, thinking about first, the issue of what is the endpoint, what is it that we're trying to show? There were many great presentations.

I was especially influenced, impressed, as I listened to John Powers, as I listened to Daniel Musher, presenting insights about this issue. My sense about this is there is not a single answer to this. There are a number of approaches that would make sense.

I am motivated, though, very much by the ICH guideline principle, stating that

the clinical endpoint preferably should be the most clinically relevant and convincing evidence and it should be a valid and reliable measure.

There are a number of clinically relevant measures. There's mortality, there are complications, there's days in the ICU, in the hospital, PROs. Certainly, as a patient, among those measures that is most profound in their benefit would be mortality, and I think there is a margin that can be defined for mortality. I'm very impressed with the evidence given, in particular by Mary Singer, about what we know about effects on mortality.

In particular, if we are conducting the study in a high-risk population, in a risk of patients that are older, that have comorbidities, if in fact you could identify such a population that would have something on the order of a 15 percent mortality in the control arm, receiving appropriate control therapy, then a margin of 10 percent I think

can be justified, even in the context of stating what would be a clinically-acceptable loss of efficacy, although that is a relative 67 percent increase in mortality that one would be essentially ruling out.

If there was a lower mortality that was achieved in that study, the issue of course is how much lower can it be and still be able to interpret the data that Mary had given as establishing the level of benefit? Obviously as well, I believe if the mortality was 10 percent, then on the same principle, allowing a 67 percent relative increase, the margin would then be 6.7 percent.

In that light, we're talking sample sizes of 500, if you could get a 15 percent baseline. It would be more toward 800, if you had a 10 percent baseline.

Some of the benefits of a mortality endpoint would be, first of all, obviously, it is very clinically relevant, very much the most significant benefit that's being provided

to patients, and would be assuring that new therapies wouldn't be meaningfully losing on that most important benefit.

Helen Boucher gave a great presentation on the issues of blinding, and one of the benefits of mortality is, as her presentation pointed out, even with best attempts, there are lots of complicated issues you have to face. Toxicity side effects, other intended effects of the treatments could lead to inadvertent risks of unblinding, and the more concrete objective that endpoint is, such as mortality, the more robust your results would be, if those types of events occurred or those types of circumstances occur.

One other quick thought, and that is in other disease areas it's not uncommon to do mega patient trials in many areas, and one of the things that makes that achievable is a concept of large simple trials. You're not needing to assess everything under the sun on

every patient in a mega patient trial.

So if this trial is very large, in the context of what we typically do in this setting, partly because we'd be doing an appropriate non-inferiority margin on a survival endpoint, it is possible to in fact look at survival on all of the people that are enrolled, but in selected sub sites, to be doing more intensive assessments for other secondary endpoints that you'd be wanting to assess.

So that it is possible to do a more cost-efficient trial, even though it's large, when you have an endpoint such as mortality.

So, in closing, there are other endpoints that could be used, and we've heard about some of them. Some of them could be composite endpoints. Some of them could be looking at time, too, that would enhance sensitivity. The issue with those, though, is it becomes more complicated to define what a margin would be for those measures, without

the same kind of evidence that we have from Mary about what the effects historically are on survival as an endpoint.

DR. GILBERT: Ed.

DR. COX: I want to start out by thanking everyone for really a very valuable two days, and a lot of great discussion on community-acquired pneumonia, and, you know, the goal of the workshop was to advance our thinking on this and to develop it further, and clearly it's done that, and I think that's been really wonderful.

You know, clearly what we're after is, you know, to try and do more informative trials in community-acquired pneumonia. I think one of the things that we heard, both yesterday and then also today, is the importance of understanding severity and defining the target population for these studies.

And, you know, from what we've heard from the presentation, you know, from

Mary Singer, you know, in order to be able to anchor what we know about treatment effect, we'll need to take into consideration, you know, the population that was studied. You know, in those studies, the type of disease that they had and how endpoints were measured, and that, in essence, will provide us with an assessment of treatment effect.

The historical data and the treatment effect shown, you know, for the patients studied in those populations look like they do provide for, you know, meaningful treatment effect that will allow for study of those diseases. So I think that's, you know, very encouraging.

Some other thoughts, just in general. I mean, one of the things that's come up over the course of the discussion too has been--and it's also rooted in the historical data--is, you know, are there strategies that could be used to enrich the patient population that was studied.

And the only other thought--and I'll just sort of throw this out there for food for thought--is as I thought about this problem, and certainly, you know, many folks have thought about this a lot, you know, one of the questions that, you know, I've sort of searched for data to help me understand things more, and that's the issue of progression. And it's something that, you know, I can say from, you know, from what I've looked at, I mean I haven't really been able to get a good feel for progression because one of the things that we see is we see a patient at a specific point in time and the question would be, you know, where would that patient be some time down the road, and, you know, some Fine class threes may, you know, turn into Fine class ones down the road, or vice-versa.

So it's something I've thought about some, and I welcome folks to think about it more, but, you know, the issue of, you know, severity and what would happen to a

patient over time, you know, perhaps somebody with bacteremia, some with pneumococcal disease.

It's an area where I think again, you know, while we're searching for data in a variety of different areas here, that seems to also be one of the areas where we wish we had more data and something that might help us to further evaluate studies and study designs. And with that I'll close. Thanks.

DR. GILBERT: Dr. Temple.

DR. TEMPLE: The main issue I think is what kind of study can we do in pneumonia that would provide persuasive evidence that a drug works.

And the only thing I've heard so far is that you could do a trial in a population like those that were studied in the past, in a population similar to the population that was studied in the past, using an endpoint similar to what was studied in the past, and we think there's probably a large

enough effect to do a non-inferiority study.

Several people have talked about evolving the population, getting less sick people, and stuff like that. You don't have any data on what the effect is in those people, and I don't hear how that can be done based on a non-inferiority design. So I think everybody has to get over that.

It's got to be a population pretty similar to that, and if you do the trial, and the mortality's 2 percent, I'm not sure you know what you've got anymore because you don't have any data on that population. So I think we've got to be thinking about getting quite sick people, probably people with pneumococcus, and it's perfectly all right if you discover that they have the disease after they're entered into the trial, but you make the group with pneumococcus your primary endpoint because that's what you've got historical data on, and outcome data on.

And I think it's going to be much

more difficult. That doesn't mean you can't look at the population that has something else. You can do that too as secondary endpoints, and I'm very enthusiastic about looking at time to progression, once you've established that you've met the non-inferiority standard. Or even without it.

I mean, if one drug had an advantage over another one on time to progression, you might find that persuasive enough--or time to improvement--you might find that persuasive enough on its own to be a basis for approval, and that would be okay too. But it's hard to count on such things.

I don't understand here, for example, how one can be talking about a composite endpoint, unless that was the endpoint that was in the studies Mary looked at. There's no way to do that. We don't have an M1 for those. Not that you wouldn't like one. But we don't have one.

The other thing, however, that

seems intriguing to me is the PK-PD, or the blood level data. You know, I'm not in this business but my guess is that penicillin and cephalosporin and things are used at huge doses because they're pretty benign, so you're not going to see things like that. But for more toxic drugs, you might be close to the level at which the effect might start to dissipate, and that means that some fraction of the population might have a dose so low, it doesn't really work in them.

And that is intriguing. We're seeing that all over. I understand Tom's reservations about whether the people with the low dose, low blood levels who do badly, have some other factor that makes them do badly. But analytic approaches maybe can resolve that. That's not a guarantee, but that seems an intriguing way to look further at maybe the less-ill populations, and I think that should be pursued, and it means to me always get a blood level in all the patients, do population

pharmacokinetics, get some idea, and maybe sharing of data by a lot of companies could help put these sorts of things on the map.

So that seems very worthwhile. But the thing that's most solid so far is these old studies with the old populations, with high mortality, and it makes it seem like in a very sick population, you really could do a non-inferiority study.

And as I said before, maybe that's enough to know that this drug works in the lung, and maybe that's what you really do know, and then people make the best of it. But that's something that advisory committees and stuff have to decide. And I think that's all I wanted to say.

DR. GILBERT: Thank you.

Robert.

DR. O'NEILL: Yes. I don't want to repeat a lot of what's said. We were talking at the break about if you wanted to do these trials and get over the comment about

everybody being on a single dose within four hours, and that problem, I think we're talking about changing the protocol in the emergency room, so that you could directly randomize folks and not have the single dose problem.

I think if you can fix that, and sort of get people on board to have a protocol where you can get randomization at the ER time, that'll really help this problem, I think, certainly in the very sick folks.

I think the other issue that struck me is how steep the mortality curve is as a function of age. So you're probably looking at 70 year olds, to have a lot of action in these trials.

A couple other practical issues. I must say that the practicality of blinding these trials, I was struck by your presentation. So whether they have to be blind, practically speaking, I don't know, but there are some issues here, particularly as Dr. Mandell's presentation this morning about

arguing for site randomization because of so many different prognostic factors that are on the ground in terms of medical culture approaches to the problem.

And I think this is a real issue. Karen Higgins reported in the mild cases of CAP. All those studies yesterday, half of them were done outside the United States. So geographic location is an issue, and I think people are ducking it, and I think we need to think about what are the implications of the design of these studies in terms of where the variability is coming from, and how that's accounted for in these trials, particularly in the non-inferiority objective, because that's a problem. That's a source of noise, and it's a source of heterogeneity, that we probably need to understand, if the prevalence of the conditions are dramatically different in the geographic areas.

We're going through this problem outside of non-inferiority in terms of

differences inside and outside the United States in multiregional studies. But that's another discussion.

With regard to the endpoints, I think Bob's correct. I think the composite endpoints are probably really only worth doing in a superiority, show-of-difference trial, not in a non-inferiority trial, because you don't know how to come up with a margin. But if you were to go down that route, I think you want to be real clear on putting composites that are actually going to move with treatment. It doesn't help you to throw an endpoint that has lousy specificity and sensitivity, because, again, it's going to get in your way, I think, in interpreting this.

And I know part of the problem, even in cardiorenal, is if you have a composite time to event, it is driven by the time to the earliest occurrence of whatever you throw into the pot.

So if hospitalization is driving

it, it's going to be driven by that, and it's really going to be driven by that in a geographic sense, if medical cultures have different ways of treating people, and it's not a real hard endpoint.

So I think it's really worth thinking about that because other areas have experienced that frustration.

And I was just curious why--I don't know how many of these trials have data monitoring committees but we did discuss yesterday the value to a data monitoring committee. But they're probably not large enough. But I'm thinking that, you know, put that back on the table in terms of the value of a data monitoring committee.

This is coming up in the context of adaptive designs, not that I'm encouraging adaptive designs in this area. But the idea is if you had a study that took a year to do, and there was some stage where you had prespecified modifications to the design,

prespecified modifications that were legitimate, and they might be something that would be adding to a composite, and one of the components that you really had pretty good information on, it's very high sensitivity and specificity, that it was moving in response to treatment.

It's just a wild idea but I'm just suggesting that in terms of using as another design feature, a data monitoring committee that might help some adaptation of the trial.

And the reason why I'm saying this is I think you're still going to be stuck with the current diagnostics available, with the mixture population that you've been dealt. Because essentially, if you can't enter people really with the disease, and you've got a mixture, and you're hoping to enrich it with as many folks that are responsive to the therapy, but if you've still got a mixed population, maybe you can somehow adapt as you go along to enrich that. But I don't know

whether that's possible or not.

DR. GILBERT: Thank you, Robert.

John.

DR. BARTLETT: I don't have a lot to add. In terms of your first question, severity, I'm not sure we can do better than a PSI score. It is a bit heavy on the age part of it. For the severity of illness, I think the PSI score is as good as we're going to be able to do, cause it's been so well studied and verified. But I do worry about what Lionel pointed out, which is it's so heavily driven by age.

In terms of the non-inferiority issue, I can't add to what's been said. In terms of the severe pneumonia issue, I think that's a real problem. People have said that's where the money is and that's where we should go. I can't imagine doing a severe pneumonia antibiotic trial.

Dale Bratzler said, well, they're taken out of the measure, but that means--you

know, look at what the consent has. I mean, these big long consents, you're not going to change that, and in the consent you're going to say, Mr. Jones, if you decide to do this, it'll delay therapy and that'll increase your risk of death.

And if you give one antibiotic, one dose of an antibiotic, that kills the study on the basis of what we've learned.

So I think it's going to be very hard to do the severe pneumonia protocol.

I also think in the contemporary situation, I think it's going to be very hard to do microbiology. Dale didn't present it but the Medicare recovery rate of a pathogen, the going rate in American hospitals right now for CAP is 11 percent. It's 5 percent for bacteremia and 6 percent for sputum bacteriology. I mean it's awful. And that's the environment in which most of us work unless we can somehow change that.

So where's the optimism in all of

this? So I would be very encouraged by some more, I think better ways to mark response. The PRO I thought was great, but we've talked about that biologic markers may work. I loved the quantitative molecular diagnostics for bacteriology and qualitative for everything else. I think that's where there may be some real good opportunities in the future.

The one other thing I would say is we've got to be mindful of harm, and that's in the way of resistance and in side effects, and the current guidelines are very heavily pointed toward C. difficile as a major complication.

In fact you can't treat community-acquired pneumonia in American hospitals without that risk, and that's being high, so that ought to be an important priority.

And then finally I'll just mention the IDSA of course wants good science, we want to reduce harm, we want to reduce abuse, and we want a lot of antibiotics. And that may

get us there if we have some way to identify pneumonia, major infectious disease, cause of death on Earth, has a high priority through the STAR Act or the TB Alliance equivalent, or something like that.

And then maybe could have a network comparable to the National Mycosis Study Group with a whole bunch of really good scientists that are doing clinical trials in sync with some science.

And it'd be lovely to have some of the NIH money go towards some of the things we've talked about; not clinical trials, but in terms of looking at the surrogate endpoints or the diagnostic markers that we're talking about, or the methods to validate the PRO.

DR. GILBERT: Very good.

Glenn.

DR. TILLOTSON: Thank you. I'm not going to expand upon any of what I would call the scientific issues that have been expounded.

One of the things I'd like to suggest, though, is that a lot of these ideals and ideas and concepts are great, but the industry's needing some answers relatively soon, please, and I don't quite know what the timelines are for these sorts of processes. But, clearly, the longer the clock ticks, the more the interest from the pharma side is diminishing.

But to put my scientific hat on briefly, one of the things--I think it was Lionel, earlier on, mentioned the CURB, CURB-65. I don't know whether it's feasible or whether it's just been a very long couple of days, good days. But can we use the CURB-65 in reverse, on a time basis? If you can use CURB-65 or CURB to identify the severity of disease and where you treat someone, can we use the reverse of that to actually determine when somebody has actually improved on a day by day, or every 12 hours, or something like that?

Clearly, those are parameters that we feel are good enough to say how sick someone is.

So can we say how good those parameters are in terms of how well they are? Maybe confusion and so forth could actually form the fundament of something that has already been tested by the BTS and quite a few other learned groups.

I don't think the Fine score could work in reverse but maybe CURB could. I don't know. And other than that, I thoroughly enjoyed participating in these couple days. Thank you.

DR. GILBERT: Thank you.

Helen.

DR. BOUCHER: I know I'm between everybody's departure. I'll be quick. Just a couple of thoughts. One struggle that I'm still having is what I feel a really competing priority, is between the need to have trials that mimic our real life. When I go to see a

patient who has pneumonia, in the emergency room, or they come to my office, I want to know what to do for that patient.

At the same time, I want to know how well the new drug works against strep pneumo, and those priorities are tough, because a lot of what we've talked about makes me convinced that the ITT is the way to go. How do we treat the patient before us? Well, we're not likely to know what they have, and I fully agree with Dr. Mandell, that although I'm excited about the possibility of new diagnostics, I just think it's going to be a while, like five years, you know, before we're really at the place where we can do that operationally.

So we're still going to be left with what we have now, which is not perfect, as others have said. So I worry about the emphasis on the strep pneumo population and that's a post-randomization event. There's no protection from randomization, about who ends

up there.

You know, people monitor to make sure there's enough, but we don't know they're the same kind of patients in each group, and you could have a false--you could see some funny thing in the comparator group that's completely random, and that worries me.

In terms of endpoints, I understand the scientific reasons for mortality but I still am troubled about this, the fact that we know from lots and lots of studies, that the dead bodies, if you line them up, they're very similar in each arm of all these studies. The drugs work, and so we're biasing again towards not seeing the difference.

So I tend to like the idea of being alive and being better in some fashion, whatever that is. If there's a quantifiable way that's on a scale, that's great.

And then finally, another comment that I think is worth offering, as something to think about as you plan trials, and that's

a formal failure analysis. In non-inferiority trials, I think we learn as much about how people fail, why they fail, and the only way that's going to be instructive is if you think about that up front, and you collect all the data you need to understand as much as we can about the deaths, about the complications that develop, about other infections that might develop. Very important, no one's mentioned that, but are people getting more fungal infections, more C. diff., more Gram-negative, you know, more other things?

Can we ascertain reasons why people had a longer length of stay, more time in the ICU? And then the defervescence issue is very important but I think one confounder that's very important to address in our protocols is what about anti-pyretics and how are we going to really get into that, and those of us who've done the fungal thing have spent a lot of time on that. So that's important.

And then finally, in the big

picture, as I listen I'm more and more convinced that we have to go towards studying older, sicker patients to get a good answer. But then, me, the practitioner who takes care of 45-year-olds, and 35-year-old HIV patients, is going to be left to figure out how to extrapolate from this data, where the risk-benefit is probably different, to my patient who's forty, and that's something I think--I hope we don't give up on studying that group. So thanks.

DR. GILBERT: Thank you. I want to thank everybody, the audience, for your attention and enthusiastic participation, the panel members, each of whom went to obviously incredible trouble to present quality information, and FDA, and certainly my co-chairs.

Any last words, sir?

DR. FLEMING: And you as well, sir.

DR. GILBERT: Thank you.

DR. FLEMING: And in anticipation

also of your added efforts on behalf of the publication.

DR. GILBERT: Thank you.

DR. COX: And I just want to share my thanks. I appreciate, you know, all the work that many folks did, the panelists, the speakers, folks involved in preparing and setting up for the meeting.

We greatly appreciate all your efforts, and thank you all for attending, and your interest and your comments. It's greatly appreciated.

[Whereupon, at 4:40 p.m., the workshop was concluded]