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

+ + + + +

THURSDAY,
JANUARY 17, 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:
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
FREDERICK NOLTE, PhD, D(ABMM), F(AAM), Professor of Pathology and Laboratory Medicine, and Director of Clinical Laboratories, Medical University of South Carolina, Charleston
MICHAEL NIEDERMAN, MD, Chairman, Department of Medicine, Winthrop-University Hospital Professor of Medicine, and Vice-Chairman
MICHAEL FINE, MD, MSc, Director, Center for Health Equity Research and Promotion, VA Pittsburgh Healthcare System
TIMOTHY F. MURPHY, MD, UB Distinguished Professor, Departments of Medicine and Microbiology & Chief of Infectious Diseases, University at Buffalo, State University of New York
KAREN HIGGINS, ScD, Statistical Team Leader, Division of Biometrics IV, CDER, FDA
THOMAS M. FILE, JR, MD, MSc, MACP, FIDSA, FCCP, Professor, Internal Medicine and Head, Infectious Disease Service, Northeastern Ohio Universities College of Medicine, and Chief, Infectious Disease Service, Summa Health System
ROGER M. ECHOLS, MD, Chief Medical Officer, Replidyne, Inc.
BRUCE PSATY, MD, PhD, Professor, Medicine, Epidemiology, and Health Services, Cardiovascular Health Research Unit, University of Washington

GEORGE H. TALBOT, MD, President, Talbot
Advisors, LLC

OTHER PARTICIPANTS:
ROBERT TEMPLE, MD, Associate Director for
Medical Policy, FDA
ROBERT O'NEILL, MD, FDA
JOHN S. BRADLEY, MD, FIDSA
KATHERINE A. LAESSIG, MD, Deputy Director,
Division of Anti-Infective and Ophthalmology
Drug Products, CDER, FDA
DENNIS DIXON, PhD, NIH
TATIANA OUSSOVA, MD, MPH, Medical Officer,
Division of Anti-Infective and Ophthalmology
Drug Products, CDER, FDA

TABLE OF CONTENTS

TABLE OF CONTENTS
(Afternoon Session)

P-R-O-C-E-E-D-I-N-G-S

(8:01 a.m.)

DR. GILBERT: I believe it's time to get started. The co-convenors are very anxious that we remain on time today in order to cover the subject. We have a lot of meaty presentations, and we want to be sure there's plenty of time for discussion, so we're going to try to strict -- to stay strictly to the time schedule as listed.

I don't think we have to emphasize to this audience the importance of community-acquired pneumonia as a major cause of morbidity and mortality. We have less than perfect therapies for pneumococcal pneumonia, necrotizing pneumonia from community-acquired MRSA and multi-drug resistant gram-negative bacilli, so clearly there is a need to ensure ongoing discovery and development of antibacterials for community-acquired pneumonia.

It's long been a dream, at least my

personal dream, to get colleagues from industry, academia, and the FDA together in the same room to collectively create solutions for a mutual problem, and here we are, and so I'm most excited about what's going to transpire over the next couple of days.

I think all of us want a regulatory system that efficiently evaluates new drugs in a fair, balanced, and clinically relevant manner so that we can ensure licensure of safe and effective drugs that will meet the medical needs of patients and their physicians.

Obviously, a lot of people to give thanks to. This meeting came together very quickly thanks to Ed Cox and colleagues at the FDA. I want to thank them for their leadership, financial support, and forbearance, especially with the neophyte that I am dealing with the bureaucracy and not knowing what I was doing half the time. Ed has been very helpful.

Industry colleagues deserve thanks

for participating and providing funding for publication of the proceedings of this meeting. The Clinical Infectious Disease has agreed to publish the proceedings. We want to thank the IDSA, who continues to recognize and support the need to facilitate the discovery, development, and licensure of new antibacterials.

So why is this such a challenge? Well, I think the operative word has been uncertainty, uncertainty in diagnosis -- only in a small percentage of the patients with current technology do we know the etiology of what we're treating -- uncertainty as to those endpoints that document a treatment effect, uncertainty as to the trial design that represents the gold standard.

Fortunately, with the faculty, the presenters that we have, I think we can address all of these issues, and the hope is that within the next two days we'll have a greater mutual understanding and hopefully

less uncertainty.

So a few housekeeping items. Strict adherence to scheduled speaking times, and to expedite that, the introduction of the speakers and the panelists, although they're all wonderful and world renowned, is going to be very, very brief. Copies of all their slides are included in the packet that you received.

In addition, if the audience wants to address questions, we have the microphones, of course, but also feel free to write out your questions on a notepad and pass them to the front. In addition, obviously, there will be time for interaction at breaks and during lunch.

We also changed the order a bit, so the program will have the safety section before the panel, so the panel will be the last hour of the day, and I want to warn the panelists, everybody in the horseshoe here, that we will ask you to distill within three

minutes your thoughts about all of the presentations that you are hearing during the day today. The idea is to get out everybody's thoughts and not miss any of the accumulated knowledge in the room.

And with that, I think I will turn this over to Ed Cox from the FDA and Tom Fleming.

DR. FLEMING: I'd like to add my welcome and to begin by thanking also my co-chairs, Ed Cox and Dave Gilbert, and with this mountain scenario here, I don't think Dave pointed out that this is the view from Ed in Washington, D.C., looking to the west, and there is Dave Gilbert there hiding behind Mt. Hood in Oregon, and the purple is me hiding behind Mt. Ranier in Washington State. So we are delighted to be here, to have all of you here.

There are many significant questions that we're going to be addressing here, and among the challenges that we're

going to be facing as we look at optimal ways to design and conduct trials in CAP will be the issue of defining the disease and the indication, defining eligible subjects for a CAP trial, but, importantly, understanding the optimal way to design and conduct trials to provide reliable evidence, not just about the efficacy of an intervention but also about its safety to be able to empower us to understand the benefit-to-risk profile.

A couple of issues that will be significantly important. One of them is, as is the case in any trial, are what are the endpoints. What are the best endpoints to be using, particularly in a definitive trial or a registrational trial.

We certainly have many key endpoints that are radiological, microbiological, and other laboratory endpoints, and these measures are certainly very important for defining the disease, defining the population, assessing prognosis,

understanding what might be the optimal group of patients to be studied in a trial, but ultimately, as we're looking at a definitive study, we need to address what it is that a patient most cares about.

Patients take therapies to provide or to obtain tangible benefit. Measures, therefore, that unequivocally reflect tangible benefit are clinical efficacy measures such as reduction in mortality risk, resolution of symptoms such as shortness of breath or a cough and prevention of clinical complications such as other infections, meningitis, et cetera.

Well, these measures on proof of concept are correlated with these clinical efficacy measures and hence do provide proof of concept, and yet correlation is not enough to be able to say that an effect on a microbiological measure will reliably indicate whether we obtain such tangible benefit, partly because there are many mechanisms of

action beyond, for example, a microbiological.

There are immuno-suppressive issues. There are many other issues that influence outcome, and an intervention could have many effects beyond those that are intended, and so the ultimate question is, as we look to obtaining reliable evidence of clinical benefit, are these the measures that we need to use, and if not, what is the scientific justification for using other measures?

Another set of challenges will be what's the control regimen as we're looking at studying a new antimicrobial intervention. Can we or should we be using placebo controls, or should we be using active controls?

And if we're using an active control, does it need to be a superiority trial, or could a non-inferiority trial be done? And a critical challenge in doing a non-inferiority trial is understanding what would be a scientifically valid margin, and

then quality of study conduct issues also impact reliability of trials, issues that relate to enrollment, adherence, and retention.

So there are many challenges that we face. There are many significant questions. To be really useful, for this conference to be really useful to the scientific community and to the regulatory community, the objective of this workshop should be more than just providing opinions about what the answers are to these questions. The objective really should be to put forward the scientific insights that will be critical to providing the enlightenment about what are these and defending what are the proper answers to these questions.

Ed.

DR. COX: Thank you, Tom. Good morning, and welcome, everybody. We're very pleased to have so many folks here today, both speakers, panelists, and also the audience in

attendance, and I just wanted to make a few comments to add sort of an additional perspective to a lot of what's already been said and why this is so challenging.

And as I think about it, and I think back to, you know, the discovery and the initial clinical use of antibacterial drugs, I mean, it happened many years ago. No question it was a major advance. It led, you know, to a situation where you were able to effectively treat infections and prevent, you know, real morbidity and mortality.

You know, this led to the incorporation of antibacterial therapy really into clinical practice, but it was really before we had more sophisticated clinical trial designs and an understanding of clinical effects and other ways of looking at clinical trials and information such that we could have quantitative interpretation of the information such that, you know, the information that we need today, where we're really trying to

understand the quantitative effect of antibacterial therapies in community-acquired pneumonia, and you'll notice this as we look at some of the data that we have, what we'll be presenting.

Mary Singer will be presenting some of the data from the historical studies of community-acquired pneumonia, and there are some real challenges looking at that information and how that correlates to what it is that we're studying today.

And, of course, it's very important that we understand treatment effect in community-acquired pneumonia trials because this allows us to design ethical, safe, and informative clinical trials, something we all want.

And then just a comment or two about the workshop. The workshop really is an opportunity for us to hear data and viewpoints really on this topic of community-acquired pneumonia and clinical trial design in

community-acquired pneumonia from a number of different folks.

It's a good opportunity for us to discuss the available science and to develop thoughts on what we know and what we don't know about community-acquired pneumonia and treatment effect, and as I've mentioned, understanding treatment effect is critical to our designing safe and informative clinical trials.

We really look forward to all the discussions, and I think a very rich discussion will take place over the next couple of days as we work through this information.

I wanted to also mention some of the differences in a workshop and an advisory committee. While we'll be talking, you know, over the next couple of days to develop the science and to hear various different viewpoints and such on community-acquired pneumonia, it really is an advisory committee

which is the more formal venue for regulatory advice that we can consider in Agency decision making.

And some of you may have noticed it posted publicly yesterday, and it's publishing in today's *Federal Register*. There is an announcement about a community-acquired pneumonia clinical trial design advisory committee that will take place on April 1 and 2, so that will follow and provide the more formal opportunity for regulatory advice.

And then just a couple of housekeeping issues. We do for speakers have a form that we're going to ask if speakers are willing and consent to sign. We plan to post the slides for the workshop on a website, so that's available out at the table.

And also so that folks know, we will be recording the session, both with a transcriptionist and then also will be making an audio recording and a webcast, so just that folks are aware of that.

And then lastly I would also share Tom and Dave's thanks to everybody. I want to thank folks at the IDSA. I want to thank my co-chairs, Dave and Tom, for all their input and hard work throughout the time that we've been planning this and for their participation today.

I want to thank all the speakers and panelists for joining us here today. There's been a number of FDA staff to work tirelessly to try and pull this together in relatively short order, and to them I thank them very much for all their work and all the other folks who made this possible, and with that I'll close and turn it over to Dave. Thank you.

DR. FLEMING: Thanks, Ed. So we've asked -- to begin the workshop we've asked John Powers to give an opening presentation on how current and emerging science can improve clinical trials of antibacterials designed to determine safety and efficacy in the treatment

of community-acquired pneumonia, so I'd like to then begin by welcoming and introducing John Powers from NIH and from the University of Maryland School of Medicine and George Washington University School of Medicine.

John?

DR. POWERS: Thanks, Tom. So in the interest of time here I'm going to address a number of topics, and one of the things I'd like to address today is that this is far more than about just non-inferiority trials and how to pick a margin.

There is a number of things that go into appropriate trial design, and what I'm going to try to do in this talk is just touch upon them briefly. I've put a lot of information in these slides, mostly as placeholders, since these slides are going to be put up on the website later, and I'm just going to really touch upon these things.

I am a consultant for a number of companies, and we were asked to put

disclosures up. I'll let you read that for yourself.

I'd like to go through four points very quickly. How did we get to the point where we are today and take off from what Ed was saying about historical evidence. Where do we want to go with clinical trials?

What are the scientific standards for evaluating safety and effectiveness, which luckily also happen to be the regulatory standards, as well? And how can we do better to address these issues?

It's very interesting that, as Ed points out, a lot of this evidence in use of antibiotics comes before the era of appropriate clinical trials, yet it was the study of infectious diseases that led those changes in adopting methodology in clinical trials.

ID trials were the first to use concurrent control groups, the first to use a placebo group, the first to use blinded

assessments of outcomes, the first to use a rudimentary randomization method of alternation, and the first to use a random sequence of numbers, just like we do today.

And I put all this information on the slide so you can read it, but it's interesting to me that we come from a very rich history in infectious diseases, but one of the issues is that even at the time that those advances were introduced, they were vigorously opposed by some clinicians. Randomization was opposed as something that inhibited the doctor's ability to make the choice for the patient.

So what that really ends up doing is confusing two very important concepts of clinical practice and clinical research. In the Belmont Report, which was published as part of the National Research Act in 1979, the very first part of that report makes a clear distinction between clinical practice, which are interventions that we choose to help an

individual, versus clinical research, which is an activity designed to test a hypothesis in groups of people to develop generalizable knowledge. Therefore, a clinical trial that's not appropriately designed cannot develop generalizable knowledge.

So what's happened, as Ed already pointed out, is that unconfirmed data becomes a part of treatment guidelines, and I found it very interesting that people always refer to treatment guidelines when they talk about clinical trials.

Treatment guidelines are about clinical practice, not about clinical research, but then what happens is that any claim about a study which attempts to confirm those data is considered unethical, because the treatment guidelines may mention it.

However, when you think about it the other way, we actually have an ethical obligation to confirm those hypotheses to evaluate whether we're actually doing more

harm than good for people, and John Ioannidis wrote a paper in *PLoS Medicine* two years ago, which actually claimed that most research findings are false, because for many scientific fields, claimed research findings may often be simply accurate measures of prevailing biases. In other words, we're studying what we think we know, rather than trying to answer the questions that remain unclear.

So in ID, there have been several assumptions whose validity is actually kind of questionable, which is what we're here to talk about today. One is that there is large treatment effects with antibiotics across all diseases, regardless of populations and severity of illness, and one of the big keys, obviously, that we're going to talk a lot about is misunderstandings about the goals and design of non-inferiority trials as a basis for evidence.

The other issue has been the thing

that we do in clinical practice a lot, and that is our reliance on in vitro testing to make appropriate choices for patients, but a clinical trial is actually designed to try to figure out whether the hypothesis that we get about in vitro activity actually translates into meaningful benefits for patients.

The other issue as we talk about these things today is that reassessment of data and quantifiable analysis are an integral part of science, so it's always good to go back and say, "Is what we're doing really appropriate? Is what we know, or what we think we know, really what we know based upon the evidence?"

So I'm actually going to skip over a lot of this stuff and just get to what we really need to do is come up with appropriate measures of effectiveness from adequate and well controlled trials, safety, which is based on a different standard of all methods reasonably applicable to show the drug is safe

under the conditions of use, and that means that safety is related to how the drug is actually used.

A drug may be safe for one use and not used for another. And then finally we want to try to balance those two together to evaluate the overall safety and effectiveness together.

Before we get on to talking about specific points, I wanted to mention that FDA usually, but not always, requires two studies to confirm findings, and confirmation is a part of science. Confirmation, though, is really not the same as replication, as it says in FDA's guidance on providing evidence of effectiveness in human drugs and biological products.

That means there is an opportunity here. It means that since, if we have to look at two trials in community-acquired pneumonia, we can actually ask two different questions in those trials.

For instance, we can look at one trial, which actually helps us pick the correct dose, and if done in the correct way, that trial can be one of the adequate and well controlled trials to support approval and can tell us something about appropriate dosing of the drug, which can then be used in a Phase 3 trial.

The other reason I bring this up is that when you read publications, often you see these two trials pooled together into a single trial, and it's important to remember that that still has the strength of evidence of only a single trial, and we still need to talk about where is the confirmation for that, as well.

So let's then go through these seven points. These are the seven points which are listed in FDA's regulations for an adequate and well controlled trial, but these are also -- and based on appropriate scientific criteria for how one would devise

evidence about the effect of a drug, and I'll let you read these for your own, since we're going to go through these one at a time.

So the first is that the trial needs to have a clear objective. Well, one of the things about objective is what are you actually studying. In clinical practice, we often operate on the theory of empirical therapy. We don't know what a person has, and we're going to give them a drug until we figure it out later.

In a clinical trial, that can actually be very problematic if you're not studying what you think you're studying. The other issue is that we want to be studying actually diseases that are of a similar pathophysiology and a natural history.

So, for instance, we would not pool all trials to get -- all studies together in Staph. aureus or in Streptococcus pneumoniae, because what would happen is we'd be underpowered for each of the individual

diseases. So if we did a trial in Strep. pneumo. disease, and there were a hundred people with pneumonia and two of them with meningitis, we obviously wouldn't be able to say a whole lot about meningitis, and the natural history of those diseases is different.

In this setting, what we're talking about is how do we relate typical and atypical pneumonias, and we'll talk some about that today, but it seems that the typical pneumonias of Strep. pneumo. and H. flu differ from most Mycoplasma and Chlamydia, but maybe Legionella leans much more towards being like the typical pneumonias in terms of its severity.

The issue that we're going to talk about a lot is obviously what is your goal in terms of do you want to show that a new drug is similar to an older one, or do you want to show that it's better? If we're developing drugs because the older drugs are no longer

effective because of resistance, it seems illogical to want to show that a new drug is similar to something you say doesn't work anymore.

So there are settings where obviously we're going to have to be talking about superiority, but we still want to see if a new drug is similar to an older drug for non-resistant infections, so we've got to address how to do that in an appropriate way.

If there is one thing that you should take out of this whole meeting, it's that non-inferiority trials do not tell you that a drug is as good as or equivalent to an older drug unless you actually show statistical superiority of the new drug to the old drug.

That is absolutely key, and I was just sort of scanning through the trials on pneumonia, and virtually every one of them uses these words "as effective as" or "equivalent" in their conclusions, but really

what a non-inferiority trial tries to do is rule out an amount by which a test drug is less effective than a control, and as Dr. Temple told me, it's a not too much inferior trial is the way to think about this.

All these trials rely on historical evidence. Therefore, they have the same biases as historically controlled trials. Can we use the data from the past to today? And the big issue is that protection from biases is less helpful in the setting of a non-inferiority trial.

The things that protect you in a superiority trial such as enrolling people that have viral illness that are going to get better anyway, they would result in a negative conclusion in a superiority trial but actually make two drugs look more similar in a superiority trial without really telling you anything about whether the drug is effective or not in pneumonia.

If the data is not available to

quantify the effect of the control under the conditions of the current trial, then a non-inferiority trial can't distinguish an effective from an ineffective drug, and you're going to hear a lot more about that.

The data from the early 1900s show that the use of antimicrobials in pneumonia is based on large treatment effects and decreasing all-cause mortality in severely ill older populations. I scanned this in, and it's grainy and terrible on purpose, because this shows you how old the data is that we're relying upon.

In fact, you can't even read those last two columns because it's so grainy, but what it does show is that from 1897 to 1905, the mortality rate was about 25 percent overall, and from 1922 to 1931, it was about 19 percent.

That's pretty close to the PORT studies, 27 percent in severe pneumonia, but what you see is that by age group, it differs

dramatically, that it goes from 5.2 percent in people between the ages of ten and 20, all the way up to -- that's actually 68.6 percent and 63 percent.

The other thing to notice is that depending upon where you study -- these are two places in England -- the rates do vary somewhat, so it's interesting to notice that, and this is the case today. Actually, the study from 1997 that the PORT group did shows pretty similar findings to this today, but this was very interesting I read in one of these articles.

It said the commonest form attacks those under 40 years of age. The period of life most favorable for the spontaneous recovery corresponds to the incidence of the type that's most amendable to serum therapy. It may therefore be difficult to determine in a serum-treated case what factor actually saves life.

So even in 1935, they realized that

if you study a population that gets better spontaneously, it's awfully difficult to figure out the effect of an intervention. In this case, it was serum therapy. So the issue here is that there is little evidence to quantify the effect of antimicrobials in pneumonia for less severe disease or disease caused by Mycoplasma and Chlamydia.

There is also little evidence of effect of antimicrobials on endpoints other than all-cause mortality, and Josh Metlay actually did a very good review of all the different kinds of outcomes in pneumonia and actually concludes that we don't have any evidence for anything other than mortality.

So there are other trial designs instead of non-inferiority that we should hopefully talk about today. There's dose response trials, which have been used recently. For instance, linezolid used this when studying vancomycin-resistant enterococcal infections.

There are placebo-controlled trials, which undoubtedly someone will mention today are unethical, but the question is if you don't know the effect of the control drug in that setting, then it's entirely ethical to do placebo-controlled trials, and the Declaration of Helsinki already points that out.

The other issue is you can do superiority to an active control, but you have to ask the question of what we're doing in this setting is we're exposing people to two experimental agents if we don't know what the effect of the older drug is and we don't know what the effect of the new drug is, and it's also an inefficient way to study something because the sample sizes for such trials need to be much larger than a more efficient placebo-controlled trial.

If you think those things are not clinically relevant, there's always the option of a three-arm trial that compares old drug to

new drug to placebo to be able to answer a clinically relevant question and also answer a treatment-effect question, as well.

There is also, as Dave Gilbert brought up, the issue of selection of subjects with a disease, and there's two questions hidden within this. One is how do you diagnose the disease syndrome of pneumonia, and two is how do you actually figure out what the microbiology is?

So for the disease syndrome, we have signs and symptoms, chest radiography, and the question is how does CT compare with chest x-ray. Where I work at NIH, nobody gets a chest x-ray. They get scanned from the minute they walk in the door from the top of their head to the bottom of their toes, and you find something this big on the CT on their chest, and it's unclear.

What does that mean in terms of is that really pneumonia or not? We'll talk a lot about biomarkers of inflammation and

microbiology, as well, and can some of these help increase the specificity of the diagnosis.

What do we do with subjects that have negative microbiological tests? That's actually up to half of people with community-acquired pneumonia, but does it matter? Is the clinical presentation of cough, fever, et cetera with a chest x-ray infiltrate specific enough for the disease syndrome that it doesn't matter what the microbiology is?

Some data actually indicates that up to a third of people who have negative cultures actually have pneumococcal disease, anyway. This is one study from Spain that I quoted here where they did CT-directed biopsies of people's lungs and actually did PCR for pneumococcus. Questionable whether you can do PCR on pleural fluid, et cetera, but that's what they actually showed was a third of those people had evidence of pneumococcal infections.

Signs and symptoms by themselves are non-specific to decide whether somebody has pneumonia, but they're still necessary to start off with to try to select the people who have -- who should get a chest x-ray, et cetera, and I quoted here a bunch of decision rules for pneumonia, all of which start off with signs and symptoms, even though no combination of signs and symptoms by itself is predictive of a person having pneumonia.

So how would biomarkers help us? Well, the issue with a lot of these biomarkers like procalcitonin is they were -- the reference standard for all of these studies is a chest x-ray. So they're actually being used to try to select people who should or should not get a chest x-ray, but in clinical trials, everybody gets a chest x-ray, so how is it going to help us in the setting of clinical trials remains an open question.

One of the ways to actually look at this is rather than sensitivity and

specificity, to look at likelihood ratios. I think these things are very interesting, because what you start with is a pre-test probability on the left based on the person's signs and symptoms. You then evaluate how much does this test help me in improving my post-test probability of diagnosis.

One of the most expensive things in ID clinical trials is all the microbiology that we do. If adding more tests is just going to increase the expense but isn't going to help us make a more specific diagnosis, then we have to question whether this is something that's really going to help us in clinical trials at all.

The fourth criteria is baseline comparability. How do we evaluate that people are baseline comparable? Well, that's what randomization is for, and it allows equal probability of distribution of severity of illness in each group, but clinical trials are looking at an average effect across the entire

group, so it's possible that if we pool severe and mild illness and the treatment effects are different, we're diluting out the treatment effects.

The other thing that I have heard said often is things like, "Well, I don't have to worry about clinician decision-making as an endpoint, because it's a randomized trial." Randomization only handles things that occur at baseline or before. They don't randomize, and we all know clinical decision-making is not random.

We make decisions for a reason, but luckily in pneumonia we do have appropriately validated severity classifications. The Patient Outcome Research Team or the Pneumonia Severity Index or the fine criteria, all the same thing with three different names, compare baseline variables to clinical outcomes of mortality, independent of the treatment administered, and would allow us to stratify subjects at baseline. Stratifying subjects

would decrease variability, which would increase the efficiency of trials and allow a smaller sample size by not diluting out the treatment effects.

The next issue is minimizing bias. A lot of these trials are double-blinded, but the issue is if the microbiology isn't blinded, they you may be able to figure out actually what the person is on.

So the other issue is that since culture results really aren't available for 24 to 48 hours, how do they help you, anyway, because I'm going to show you some evidence that says most people are better or on their way to getting better in 24 to 48 hours, so you can evaluate the clinical outcomes in those people.

The second thing is if we're going to evaluate what resistance in vitro actually means, we've got to blind the microbiology, because we have to compare it to a blinded assessment of how the patient is actually

doing clinically to help us better determine what resistance means, and the big issue now is what does resistance in streptococcus pneumoniae mean? There is a lot of debate as to whether what we called resistant in the past really results in worse clinical outcomes for patients.

The other issue in this regard is the issue of concomitant medications. Daptomycin was studied in the trial that's going to be published very soon where they did a post-hoc subgroup analysis that showed that people that got even one does of antibiotic prior to enrollment had a much better success rate than those who didn't.

There is also the issue of concomitant medications on therapy. If you happen to be unlucky enough to be studying a new cephalosporin today for pneumonia, some people will demand that you add a second drug to it. Otherwise, it's unethical to do so.

All of this is based upon

observational data. How would you figure out the effect of your cephalosporin if the person is getting a macrolide, as well, that overlaps in spectrum? Very difficult for you to figure that out, yet the evidence for combination therapy decreasing pneumonia is based on observational studies.

Two meta analyses in the last few years actually say there isn't a treatment effect when you pull all the evidence together, and this study by Paul is very interesting. What he did is he did propensity scoring on the people who got combination therapy versus the people who didn't.

It turns out that the people who get combination therapy are on average less sick and younger, and we already showed you the data from the 1930s on what happens with younger people. Why? Because people who are less sick is a clue to us as clinicians of a typical pneumonia, which is mycoplasma chlamydia, which are less severe.

So it could be that getting combination therapy is only a marker for less severe disease. Therefore, the reason why those people do better is really related to who they are, rather than the interventions that they receive, but this is a real problem if you're trying to develop a new therapy that's not a quinolone today, because this issue of combination therapy becomes problematic.

Endpoints is a big issue in this regard, so the clinical -- as Tom Fleming already discussed with you, clinical endpoints are direct measures of patient benefit. They would be things like mortality, is the functioning better, non-fatal clinical events like can we prevent someone from developing empyema or resolution of symptoms.

Surrogate variables, on the other hand, are defined in the glossary of the International Conference on Harmonization E9 guidance as indirect measures of clinical

benefit. ICH-E9 actually points out that you should use these when it's not feasible or practical to measure clinical outcomes, but in this disease, we can actually measure clinical outcomes in the space of a couple of days, so have to ask the question why do we need surrogate variables in an illness that has such a short time course?

The other issue is in these trials most subjects don't have microbiological data baseline that's positive, over half of them, and then when you go back to them after they feel better, even more of them you can't get follow-up cultures from.

They're not coughing anymore. They didn't have baseline positive blood cultures to start with, so it's not informative. So microbiology really doesn't help us a whole lot.

The other issue is that these trials have used a categorization of presumed eradication, and it seems rather unscientific

to presume anything in a clinical trial. You either measure it, or you don't, and it's based on the fact that if a person is better, their organism must be gone, and, actually, we don't know that, so we're presuming something that we don't know, and, actually, that makes these analyses actually even less helpful.

The other issue is combining biomarkers like body temperature, heart rate, blood pressure, O2 saturation. It doesn't turn them into clinical endpoints, nor does it increase their validity, and I bring this up because there is an article in JAMA on time to stability, which evaluates the number of endpoints, all of which are biomarkers except change in mental status and ability to eat were the two things that were actually clinical endpoints as a part of that.

It would be very helpful to develop well defined clinical outcome criteria, which is independent of clinician judgment. This is an article by Archie Cochrane in 1951, where

he took seven clinicians, Clinician A, B, C, and he pulled three of them together in D. He asked them to interview 300 coal miners. Same 300 miners each person interviewed.

Actually, they randomized them, took random groups, and they gave each of the clinicians, and they allowed the clinicians to ask these subjects, patients, about coughs, sputum, chest tightness, pain, dyspnea. Didn't tell the clinicians how to ask the questions.

Just ask them about these things, and you don't need to do statistics on these numbers to see the variability between what one clinician found and the other clinician found, right. So this gives you some pause about clinician reported outcomes if we're going to talk a lot about patient reported outcomes.

So Doctor A has got a report of 20 percent of people coughing, while Doctor C doubled that. Forty percent of the people

were coughing, so there's a wide variability if clinicians are left to their own devices and just allowed to ask questions in a non-structured way.

So the other issue is that the current endpoints in these trials are based on "enough improvement as judged by the clinician so that no further antimicrobial therapy is required." So there's a couple of issues with this.

One is improvement is not a dichotomous measure. Some people can improve a little. Some people can improve a lot, and actually the FDA guidance from 1992 on antimicrobial development actually cautions against using an end point that includes improvement.

This actually also is not a direct measure of patient benefit. What are we measuring here? We're measuring clinician decision-making, someone's judgment that the patient doesn't need any more drug, or the

subject in the trial doesn't need any more drug, and we've never really evaluated what's the inter and intra observer variability of that decision-making process.

So what would be nice would be to develop patient-reported outcome measures in symptomatic diseases, which would allow us more valid and reliable measures, and, actually, in pneumonia there already is one.

Donna Lamping has one that was published in *Chest* in 2002. Obviously, FDA would need to review that primary data on how that was developed, but at least she has it laid out as how they interviewed patients. They interviewed clinicians, they got all the information, and they actually did a structured evaluation of that PRO instrument.

The other issue is when do we measure these outcomes, and how do we measure them? Time-to-event analyses may be actually more informative, because they may increase our power to detect differences. When we look

at a fixed time point, we're looking at one point in time per patient.

If we look at a time-to-event analysis, we're gathering all the information up that point, which actually can decrease the variability, decrease the sample size, and give us more precision. The problem with this is we're going to need to measure early enough so that we'll actually be able to detect when the change actually happens.

This study by Torres used the Lamping PRO, but they only measured people at baseline, day three to five, and test of cure only. Well, when I show you some other data tomorrow, I'll actually show you some curves from a 1945 study by Max Finland and colleagues that shows the vast majority of people are better in two days from pneumonia.

So if that's the truth, then measuring a PRO only at day three to five is going to be too far out to detect any differences, anyway, so we're going to need to

measure serial measurements, actually on a very short time period. That can't be done by having a clinician interview the patient every six hours. It's just not feasible.

It can be done, though, if you use a PRO, which is on a palm hand-held or a patient diary that the person can answer at a specific time point, and the technology is such now that these things will actually ring and buzz and jump off the table and remind you when to fill these things out.

So the other issue is we need to keep in mind that if we can't see a response on a time-to-event analysis, it's unlikely that a fixed time point analysis is going to show us anything, and after reviewing some of this data from, you know, the early 1900s, what's pretty clear is that if you look at days 17 to 21, when a lot of current trials do, there's no difference to detect at that point, anyway, because most of the people have recovered if they're going to or not.

The last issue I want to touch on is appropriate analysis. One of the issues in these trials is the idea of per protocol analyses versus intent-to-treat analyses, and other people are going to address that, so I'm not going to get into that, but there is one thing that's pretty clear, no matter which one you choose, and that is that the things that we use to exclude people from the per protocol analysis, we really have to start questioning their validity.

One of the things is people haven't received enough therapy. Usually if the person hasn't received at least three days of therapy, they're considered unevaluable.

When you go back and look at Max Finland's data and realize that that's where all the treatment effect is, and that's where all the events are, it doesn't make a whole lot of sense to be excluding people who don't take three days of drug, because if you're going to fail, that's when people fail, most

commonly.

So the other issue is it also doesn't make sense from a clinical trials and statistical point of view to be excluding people post-randomization, especially if they die. If somebody dies on day two or three, that's important to know, and it also ignores the fact that they may be dying because the drug is doing harm, and mortality is the one thing that really crosses the line between safety and effectiveness that we want to look at closely.

So we need to evaluate both intent-to-treat, modified intent-to-treat, per protocol analyses, and look at all of these things to get a clearer picture of what's going on, and this is even more problematic in non-inferiority trials, which others are going to address.

We also need to address this issue of appropriate adjustments for multiple comparisons when we're looking at secondary

endpoints and subgroup analyses, and I'm going to address that tomorrow, as well as gatekeeper or hierarchical testing of hypotheses so that we can test multiple hypotheses without having to increase the sample size.

So like in the era of the first trial in infectious diseases, we have to look at these challenges as a real opportunity to answer clinically relevant questions. When Fibinger first did that trial on diphtheria toxin -- by the way, which the end point was all-cause mortality -- he did it because a previous guy in France named Roux had done a trial which actually showed huge effect of diphtheria toxin, but he had no control, no randomization, no nothing.

So what he did was he actually did a concurrent control with people that he injected with a non-diphtheria toxin, but it happened that that wintertime, nobody got diphtheria, so he had a negative trial result,

but he was smart enough to realize that, "Well, I have to do this again, because I got a negative result."

So the truth of that was you can learn a lot from negative results, too, as well, but the other issue is Fibinger was the guy who realized that just because the guy in France did his trial one way, he wasn't going to be bound by that precedent and that we can learn lessons from the past to actually help us design trials better in the future.

The other issue here is we need to address all seven of those criteria for effectiveness, as well as appropriate safety evaluations. This is not just about picking a margin and leaving everything else about these trials the same. There is a number of things that we want to look at.

If we can make these trials better, we can reach multiple goals. We can get clinically relevant answers that will help clinical practice and make decisions.

We can aid regulators make better decisions about the safety and effectiveness of these drugs, and by making some of these changes, we can increase the efficiency of clinical trials for drug sponsors and allow them to get where they want to go in a more efficient and faster way while still helping patients along the way.

So even though it looks like we have to cross this big desert to get where we're going, hopefully there is a nice, shiny star at the end here that we can actually get something that's better for everybody.

Thank you.

DR. FLEMING: Thank you very much, John. You set a great standard for the speakers here. You provided some excellent insights, and you did so coming in two minutes early.

We do have time for Q&A and look for questions for John, comments.

DR. GILBERT: John, can I ask -- I

know you're going to amplify on this subject tomorrow, and the question deals with subsets, and as we move forward with our diagnostic ability, it seems like we'll be able to get some of the noise out of the system, in particular viral versus bacterial disease.

So how does that influence some of these entry criteria and evaluation issues that you presented to us?

DR. POWERS: I'm going to talk about that tomorrow in a little more detail. It actually does two things. One is it may increase the efficiency of the trial in terms of being able to focus your treatment effect on the people who might benefit the most.

So if you have a drug for methicillin-resistant Staph aureus, and you really believe it's superior to an older drug, you can now do a superiority trial in those people by focusing on them, and FDA just approved last week a new rapid test for detecting Staph aureus in the blood. The flip

side of that is, though, you may have to screen more people to get those folks.

So there is a flip side to that, but I think it will greatly help, and if we can then actually translate those out into clinical practice, as well, we'll be able to focus on who actually most benefits from the drugs and not use inappropriate therapy, which drives resistance in the first place.

Bob?

DR. TEMPLE: If you do something like that, how do you know what margin to use? You don't -- your historical data didn't select the population that way. How do you translate?

DR. POWERS: Right. I don't think you can do a non-inferiority trial for that. What I'm talking about is suppose you want to say, "My drug is more effective than vancomycin in the treatment of pneumonia due to Staph aureus." Then you could actually design a superiority trial to test that

hypothesis and more efficiently select the people. I think doing a non-inferiority trial is going to be very tricky if you don't have the historical evidence, just like for any other non-inferiority trial.

Ed?

DR. COX: John, thanks for your presentation. You mentioned the Lamping PRO and looking at shorter term outcomes, I mean, in essence within hours of starting therapy, and I'm just curious if you have any thoughts on, you know, if somebody were to approach it using a shorter term PRO tool, looking at that very early time frame, if there are other things, too, that one might want to look at, you know, in later time frames, also, thinking about potential complications that may not develop until later points in time and some of the limitations thereof.

DR. POWERS: I'm going to show that tomorrow. There is actually data from a guy named Cecil, you know, the Cecil textbook, the

same guy, and when I show you the numbers, even in the untreated period, the number of complications is very low.

So, for instance, empyema occurs in about 6.5 percent of people, so if you wanted to show a difference on those things and say, "My antibiotic prevents emypema better than somebody else's," the sample size for that would have to be extraordinarily large.

So I'm going to talk about it tomorrow when we talk about putting things together in a composite end point. You may want to make that one part of the composite, because it is an important clinical outcome, but it won't be driving the overall outcome, because there's just too few events to be able to measure that much.

DR. FLEMING: Okay, other comments, questions? By the way, I should note that anyone in the audience who has questions or comments, I think we have a live mic there. Please feel free to use it.

DR. REX: John Rex from -- it is definitely on now. Good morning. John Rex from AstraZeneca. Thank you, John. It was a fun talk.

On one of your slides you make the observation talking about similarity versus superiority. If an older drug is no longer effective due to resistance, it seems logical to show superiority of newer drugs.

While I agree with you in principle, that comment overlooks a real time line issue in terms of bringing forward new drugs. Part of what we do in the industry is say to ourselves, "Well, I see a title wave coming of resistant gram-negatives," and the time from gleam in discovery scientist's eye to drug exiting Phase 3, you know, it's ten years or so. It's a long time, and even when I get sort of close to the end of that, you know, the rates of resistance may still only be 10 or 15 percent, so it's actually quite hard to do what you have pointed to here.

Most of what we're going to be studying will be drugs, I'm sorry, bugs that are still susceptible to the old agents, but we're trying to get out ahead of a coming problem with resistance. So I'm just asking you to talk about reconciling those two issues, because it's hard to do.

DR. POWERS: Right. I'm going to talk about it in a lot more detail tomorrow when we talk about endpoints and how to measure them, and I don't want to -- I stayed on time, and I want to stay on time, so all I'm going to say is that --

DR. FLEMING: John, just -- this is really a question that would be helpful if you gave some initial sense. I mean, it's a great example. So you have an effective antibiotic, antimicrobial. Eventually, there are issues of resistance.

There is an interest in having a more global set of options, alternatives, so if you do a non-inferiority trial against that

existing antimicrobial while it's still effective, the issue is that would establish whether it's effective, but your key question is you're also motivating the ability to have an alternative when that standard agent itself now has resistance.

So does this, in fact, give you -- does this approach give you the knowledge that you now have another therapy that will be effective in the resistant patients of the future to the standard intervention.

DR. POWERS: Right. So this comes back to, actually, what you can claim, and this is more of a question to ask FDA folks, because I don't work there anymore, but the issue is supposed you wanted to get a claim for vancomycin resistant Staph aureus. There are, what, nine people that have had that?

So at this point in time, it would be very challenging, if not impossible, to do a trial for that, but it would also be very challenging to give someone a claim for that,

because FDA has to base those analyses on what actually happens to human beings in clinical outcomes.

There seems to be what we need is a mechanism that somewhere down the line that we can go back and take older drugs and actually study them to see what their effects are in these settings, and I'll give you the example of clindamycin. When clindamycin was approved, MRSA was like a distant thing to think about, but now NIH is going back and doing a study of clindamycin and trimethoprim sulfa, in fact, compared to placebo in skin infections to see whether they have an effect or not.

The question is can there be some mechanism by which drug sponsors can take their drug back, study a new indication, and make that somehow palatable for them. That goes beyond this discussion, but I think it's very hard to say you could get an approval from FDA for something that might happen in

the future. That puts them in a pretty sticky spot. We'll talk about it tomorrow, because I'm going to bring up how to look at this in a more detailed way when we talk about outcomes.

DR. FLEMING: But just to probe on this just for a moment further, so you find an experimental antibiotic that you compare to vancomycin in the setting where the idea is eventually, when it's much more frequent that you would have vancomycin-resistant Staph aureus, you want something to use in that setting.

The question is if you do the non-inferiority trial of your new antibiotic against vancomycin in people who aren't resistant, how do you know that when you have vancomycin-resistant Staph aureus that this new experimental antibiotic will be effective in those patients? How do we know that without doing a superiority trial in patients who are vancomycin-resistance Staph aureus?

DR. REX: That's one -- I was going

to say to John you didn't quite answer the question I was putting to you, and you pointed at one half that wasn't quite pointed out, and the other half of it is that we're spending a lot of time talking about superiority versus non-inferiority.

I think it's important to recognize that it may really be technically impossible for me to do. I can't find the vancomycin-resistant Staph aureus out there, so I can't actually prove the superiority, and not only that. You don't want me to wait.

You want me to go on and get the drug studied on the market available such that when you actually need it, and for some other reason it's actually done, and so I'm not coming to say, "How do I get a label for vancomycin-resistant Staph aureus?" I'm saying, "How do I get the drug -- you know, do I have to do a superiority study when I can't actually do one?" which is kind of what your sentence implied to me.

DR. POWERS: Right. That's not what I was implying, and, again, I've got a whole section of five slides on this tomorrow, but the idea is that you could study your drug against vancomycin in an appropriately designed non-inferiority trial and claim that you had similar effect to vancomycin in disease X due to Staph aureus.

What becomes problematic is to say, "In the future, my drug will be superior to vancomycin for the treatment of vancomycin-resistant Staph aureus." Those are two different statements.

You are not barred from saying, "My drug is effective in Disease X due to Staph aureus." It's the resistance part of it that becomes more problematic, and, again, I'm not doing justice to this, because I'm speeding through it, trying to get -- Bob, you want to --

DR. TEMPLE: Well, don't -- you know, I'm not the -- I'm not an ID person, so

this is probably stupid, but if you encountered in your -- I mean, if you had a drug that was effective in pneumonia, community-acquired pneumonia, you don't test it against every conceivable organism, do you?

DR. POWERS: No.

DR. TEMPLE: I mean, you use your in vitro methods to guide you to a degree, so if you had established through the presumably non-inferiority mechanism that the drug works, this new potentially useful drug that's good in resistance, I don't -- wouldn't -- would we ask for a documentation that it works in the resistant organism all the time or not?

DR. POWERS: I think the question is really --

DR. TEMPLE: What you need to know is that it works in pneumonia.

DR. POWERS: Right. Exactly. That's what I'm getting at is that you need to know the broader question first. The real tricky part becomes you're relying on a

historical assessment that the in vitro test actually predicts failure, and the example I'm going to go through tomorrow is exactly this one.

Pneumococcal resistance in pneumonia -- we had called any organism that had an MIC greater than two was resistant. Now looking back through that, it appears that organisms with MICs of two and four, that the success rates are actually quite similar to the people who have MICs of .6 and below.

So we had, in essence, incorrectly defined resistance, and that's because we often based resistance on case series, not on actual randomized clinical trials, and that's difficult to do. I'm not saying that that's an easy thing to accomplish, but it relies on a historical assessment that the old drug is ineffective, and that's the question.

So, for instance, let's look at skin infections in MRSA. AAC published something in December that actually showed a

trial in skin abscesses that the success rate for people with just lancing the abscess was 90 percent.

And now the control drug was cephalexin, which you wouldn't expect to have much activity against MRSA, but the fact that 90 percent of people got better, and the drug couldn't show superiority, even in a resistant pathogen, starts to say, you know, maybe resistance is different at different sites in the body, too. That might not be the same for pneumonia as it is for a skin abscess. So it really gets down to the definition of what resistance is, as well, and I think that's a whole other kettle of fish.

DR. GILBERT: John, again, you'll probably address some of this tomorrow, but back to the patient reported observations, just to fine tune that a little bit, isn't that tough to do in severe disease when one of the criteria is the patient is confused, et cetera, et cetera?

DR. POWERS: Right.

DR. GILBERT: The treatment effect is early, as you pointed out, so I'm going to address, actually, the patient-reported observations in a couple hours, and I'm pretty excited about it, but I'm not terribly excited about it in severe disease.

DR. POWERS: But in severe disease, the end point is really all-cause mortality, so what I'm going to go through tomorrow is various pieces of the end point and how you pick and choose among them, depending upon the disease.

So patient-reported outcomes can be helpful in people that are awake and talking who have less severe disease, or they can be a component of a bigger end point, so not everybody that gets -- in fact, very few of the people that get admitted to the hospital with pneumonia end up in the ICU and are confused. The majority of people that get hospitalized end up on a regular ward floor,

so maybe a PRO could be a supportive outcome in those people, as well.

So in one case, it can be the primary outcome, in another case, it can be a supportive outcome, and in the third case, you wouldn't use it at all, because in obtunded people it wouldn't be useful. So this gets to actually using the right end point for the right patient population, which you're going to talk about, too, coming up soon.

DR. SPELLBERG: John, really quickly, I almost hesitate to ask this, but I think it's the elephant in the room, and I think one of the most important things we need to decide is the issue of placebo control, so I'm going to start the ball rolling by asking you do you think that there -- clearly you're correct that if you don't know if your comparator is active, it is by definition not unethical to do a placebo-controlled study, but there is a complexity in that all these patients have primary care physicians, and if

those primary care physicians won't allow those patients to be enrolled in a study because they believe that the comparator should work, how do you address that? I mean, how does that come into play?

DR. POWERS: Right. I think -- I'll expand it even one further, and if the IRB believes that, as well.

DR. SPELLBERG: Exactly, yes.

DR. POWERS: So I think the first thing is that I think this is also -- it's on us, part of this, is to explain to people what the data actually is. When I was at FDA, we got a response from an IRB that said, "We're not approving this trial in AECB, because we don't think it's ethical."

What they sent back was the abstract of a review article, so I would point to those people to the Belmont Report, where the beneficence part of it says it is incumbent upon you before you do a clinical trial to go back and actually review all of

the evidence, not a review article, actually the evidence.

The second thing is I would say there are other trial designs besides placebo control that can be used in this setting. I think we way under use those response trials. The reason why people are opposed to those is they say, "Well, I'm not randomizing somebody to a group that's less effective."

What's a non-inferiority trial? You're randomizing somebody to something that might be less effective, but it's actually the belief that non-inferiority trials show the two drugs are equal, which actually make people argue against those response trials, and actually, those response trials would help us validate some of the stuff we said about pharmacodynamic analyses, as well, and close the loop on that to see whether actually pharmacodynamics can actually predict better clinical outcomes.

We keep saying that increased

potency in the test tube means something for patients, but we haven't shown it yet, and actually dose response trials would actually allow us to be able to test that hypothesis and provide some evidence for it. So we don't always have to do placebo.

DR. FLEMING: So let's move ahead to a couple quick thoughts. Bob O'Neill I see in the back. Bob, we've got a seat for you right up here in the front, and just one real quick follow-up thought to Bob Temple's question.

Certainly it is the case that if I do a non-inferiority trial against an effective antimicrobial, and there isn't resistance, then that is evidence for benefit. The issue comes in that often we argue we can have a more lenient non-inferiority margin, because this new therapy will give us an alternative when there is resistance, and therein lies the complex issue, because you don't know that that agent will, in fact, be effective once resistance occurs.

So what we'd like to do is make a transition to Dave.

DR. GILBERT: So when the convenors were trying to figure out how to address these complicated issues, we decided to present two theoretical patients. The one for today is the patient with modest severity community-acquired pneumonia, and then tomorrow we'll move on to the patient that has more severe disease.

The patient with modest represents 80 percent of all of the patients with community-acquired pneumonia. The patients with severe disease are 20 percent, and obviously they're usually in the hospital. The patient with modest disease is usually treated as an outpatient, and the point is to get us thinking about the clinical setting that we're dealing with.

So this is not a real patient, of course. This is a 35-year-old male resident of Boston -- I'm not picking on Boston -- who

presents with fever and a cough. The patient was well until three days earlier, when he suffered the onset of nasal stuffiness, a mild sore throat, and a cough productive of only small amounts of clear secretions.

He visited his physician, motivated by a fever of 38.3 degrees. By the time he got to the physician, he had some purulent secretions and spasms of coughing.

It's March. ID doctors always want the epidemiology. He lives in the city. There is no problems with his home. There are no obvious risk factors for Legionella is the point of that.

His wife is well, but his 11-year-old child is recovering from a nagging cough that lasted ten to 14 days. All four children have been fully immunized. Of course, we have a pet parakeet, but the parakeet is well for the last five years.

There is no recent travel, but the patient does smoke a pack per day and has done

so since age 15 and admits that, especially during the winter months, when he wakes up he has to clear out his lungs and produces a few teaspoons of purulent sputum.

The rest of the history database is pretty negative, no pertinent past medical history. Patient is on no prescription medications, has no allergies, does smoke, and uses alcohol in moderation.

Exam confirms that the patient is febrile, a little tachycardic. Blood pressure is okay, maybe a very slight increase in the respiratory rate, but the oxygen saturation on room air is satisfactory.

There is some hyperemia of the nasal mucosa and erythema of the oral pharynx, no adenopathy. There are crackles heard at the right lung base, and the patient has a spasm of coughing during the exam and produces a very small plug of purulent secretion.

So, to get to the lab work, hemoglobin hematocrit are fine. White count

is slightly increased with maybe a few increased in immature polymorphonuclear leukocytes. Platelets are fine. Chemistry screen and urinalysis are normal.

Somebody's going to say, "Well, gee, my doctor doesn't do a multi-chemistry screen," but for purpose of this hypothetical patient, I threw that in, but, of course, the chest x-ray shows bilateral lower lobe infiltrates that was a bit asymmetrical, more pronounced on the right than on the left.

So if we apply the Pneumonia Severity Index, this is a Class 1. If you prefer the CURB-65 prediction score, it also gets a score of 1, making the patient a candidate for outpatient therapy.

So the doctor knew that. He ordered no micro biologic tests whatsoever and empirically prescribed a respiratory fluoroquinolone.

Against medical advice, the patient continued to smoke. The fever resolved over

three days. The cough gradually returned to his baseline pattern over the subsequent seven to 10 days.

So that raises lots of questions. Is the patient a candidate, or would the patient be a candidate for a placebo-controlled or a delayed treatment or an active control trial? We've just discussed some of the issues that are involved about that, and statistically would this be a superiority or non-inferiority trial?

What severity of illness is appropriate for inclusion in an outpatient treatment trial, the severity of illness determined by which scoring system? Is there any substantive difference between the various prognostic severity systems that we've been using?

Which diagnostic tests? Now this is in a trial setting. Which diagnostic test makes sense for virus, for typical bacteria, for atypical bacteria? And the next couple of

speakers are going to directly address the viral etiology, and tomorrow we'll hear more about modern testing to detect typical bacteria.

What's the most appropriate and valid clinical endpoints? And around 10:30 or 11:00 this morning, we're going to begin to discuss that.

And how do you blind the treatment arms in the various methods of blinding? Which makes the most sense, and which has the most powerful impact on interpretation of the results of clinical trials?

And then, obviously, the flip side. How do we monitor adverse drug effects, especially in the patients who are treated in an outpatient setting, rather than being under our direct observation in the hospital?

So that's our hypothetical patient this morning. Are there any questions on the patient population? Tom?

DR. FILE: Well, this isn't on the

patient population, but just for clarification, although it won't make a difference in site of care, there is a slight difference in the outcome, at least based on the study by Kim, et al concerning the CURB-65, but wouldn't this be a CURB-65 score of zero, unless he's confused?

DR. GILBERT: Yes, it would. I was trying to get the patient to -- my hypothetical patient -- between a zero and a one.

DR. FILE: Okay.

DR. GILBERT: And it probably would --

DR. FILE: It's not going to make a difference of site of care, but there is a little difference in outcome when you look at the -- at least mortality --

DR. GILBERT: Yes.

DR. FILE: -- because there is virtually no mortality if it's zero. There's a little bit for one.

DR. GILBERT: Yes, we're still -- you'd agree that we're still at a mortality rate of under five percent at the worst sort of setting.

Yes, George?

DR. TALBOT: I think there are a number of interesting issues that your scenarios raise, and I think it goes back a bit to what John Powers was talking about in terms of the difference between clinical practice and clinical research.

So, for example, I think the clinical scenarios, both of them, reflect the way a clinician would think, and that's useful, and that's the way the real world works. Now for clinical research, though, there are some issues hidden in there that I think should be defined and discussed up front.

For example, what really are the important components of severity? Is it PORT alone? Is it PORT plus some other

characteristics? Your scenario suggests that requiring hospitalization is another component of severity, and I think we need to question whether that's true or not for clinical trial purposes.

I think that the requiring hospitalization, again, makes intuitive sense for a clinician, but how do you define that objectively? It could be socioeconomic factors, but that may not play into severity, so, although I believe that requiring hospitalization has a pragmatic aspect, there needs to be more clarity about whether that plays in on top of PORT.

So one key question in these scenarios is how do we define severity not for clinical practice but for clinical trial purposes? And that, of course, extends then to the question of how do we define where an NI approach is appropriate, and if so, what the NI margins should be.

The other assumption that I think

is in some of the scenarios is this distinction of mild, moderate, and severe, and I think we need to discuss that in more detail, too, and how it relates to PORT or any other characteristics of severity that one could define.

My impression is that mild, moderate, severe has really been sort of a labeling thing. Severe is defined variably, but labeling and clinical practice, again, are different from clinical trials, so let's be explicit about what we mean by mild, moderate, and severe before we go on with the discussion and make a lot of base assumptions about these severity questions.

DR. GILBERT: Well, I'll ask Ed to comment and then whoever else he wishes to comment from the Agency. I thought the Agency at this point, and I may be off base, divides community-acquired pneumonia into two categories, mild/moderate on the one hand and severe on the other and hence our scenarios

and this -- well, I'll just be quiet there. So, Ed, what do you think about George's comments?

DR. COX: Sure, yes. To Dave's point, you know, typically the labels that are out there for drugs for community-acquired pneumonia, oral drugs, are typically labeled for mild to moderate pneumonia to reflect the population in which they were studied.

You know, a drug that's available in an IV formulation, and many are available both in IV and PO, typically the indication would just say community-acquired pneumonia, and there wouldn't be any limitations on the mild to moderate.

And I think, George, too, you're also getting to the issue of, you know, severity of illness and what's the best way to index that so that we have a feel for prognosis, because I guess the question that I'm thinking about, and we'll hear more about this soon, is, you know, the patient that Dave

just presented, you know, he's asking the provocative question of what type of trial design, you know, would be possible for such a patient, and I guess I'm starting to think, you know, are there things that would help us to further identify, you know, what the prognosis in this patient would be.

I mean, can we quickly identify a patient who's, you know, likely to have a good enough outcome, and it's got to be very high, given, you know, that community-acquired pneumonia can progress such that, you know, you would be willing to consider either a treatment delay or something like that, and that's a rhetorical question.

You know, is that a possibility? You know, could you do a rapid test to rule out Legionella? Could you do a rapid test to rule out strep pneumo? Could you -- you know, is that possible? And I guess we'll hear more about rapid testing and can we sort of further define this patient's severity and, you know,

his risk for a bad outcome.

DR. TALBOT: Well, just to follow up on that, I agree, and I think you mentioned that mild, moderate and severe are sort of label issues, so the thing that needs to be defined for me and for this discussion is, well, is it really appropriate to lump mild and moderate, and how, in fact, do you define that?

Maybe, if we're going to talk about superiority design, placebo control design, maybe it's mild, and then maybe moderate and severe are fine for non-inferiority. So let's not get locked into mild/moderate versus severe.

The other question is, okay, if it is mild, how do we define that objectively without the potential confounders imposed by terms like "requiring hospitalization"? So mild should be PORT 1, maybe, plus not a socioeconomic decision to hospitalize but some other objective pathophysiologic parameter

that makes that PORT 1 patient, you know, at lower risk or at higher risk, because PORT -- and Dr. Fine can comment on this. There is variability in PORT in terms of which class you're in, so it's useful to have additional factors to define severity, but they again should be evidence-based.

So my plea is let's not start out assuming it's mild plus moderate versus severe, and let's not start out thinking that we can define objectively requiring hospitalization. I would rather that these scenarios have said PORT 1, not having other pathophysiologic characteristics, versus PORT 2, 3, 4 with A, B, and C.

DR. GILBERT: Go ahead.

DR. COX: Your comment is a fair one, George, and I think that's, you know, why we're here today. I mean, let's talk about what is severity, and let's try and, you know, think about, you know, the best way to define that so that we can design clinical trials

that will be most informative and clinical trials that will be safe, and, no, I agree.

DR. GILBERT: Are there other comments? Yes, please.

DR. EISENSTEIN: Good morning. Barry Eisenstein from Cubist, a comment and a question. The comment is that although there have been discussions about dose, I haven't heard anything about duration of therapy, and John Powers has previously talked about the selection of drug resistance as a "side effect," and given that longer duration therapy, particularly when not needed, may actually produce increased resistance, one could actually be dealing with a safety issue on the other side, so I'd like to know what sort of thoughts there are about incorporating duration as well as dose.

And then the other question that my colleague Bill Martone has asked me to raise, going along with George's comments about mild versus moderate, if one is able to get

approval for a trial based on a moderate cap, does that enable the manufacturer to also claim efficacy in mild infection?

I know that may sound a little bit backwards, but perhaps if mild is going to get better on its own anyway and needs a placebo-controlled trial to show superiority and you don't have that, but you do have efficacy in moderate, how do you then enable the sales force to talk about mild pneumonia?

DR. GILBERT: Well, on the first point, which was duration, obviously that would be protocol-defined, and it's a valid and very important point. Hopefully, we'll have further discussion on that as we proceed, and then I'll have Ed respond about if you get a label for moderate can you therefore claim efficacy for mild.

DR. COX: You know, I think, Barry, you know, some of your comments in your question, I mean, I think really get at the key issue here. You know, generally, I mean,

if you work, and you work in a severe condition, that provides, you know, really good evidence that the drug works, and it becomes helpful and supportive information to help you when you're looking at less severe disease, but I think you've also in your question anticipated one of the issues here, which is, you know, can you say it works in the less severe disease, and that's dependent upon knowing that, in fact, there is an effect there.

So while it is, you know, very helpful to have clear evidence of efficacy, you know, implicit in your question was is there a treatment effect in that mild disease, because, you know, it starts to, you know, get at a very sort of difficult issue of, you know, are we extrapolating down to something where, you know, it's unclear that there is an effect.

But, you know, that said, you know, clearly if you showed an effect in moderate

disease, well, then that would be an important thing, and that would be very helpful, because that would be a population where you could, we would hope, and that's what we're here all talking about today, show a treatment effect, and if you showed your drug to be safe and effective in moderate community-acquired pneumonia, you'd have something.

DR. FLEMING: I think this question here brings out one of the key aspects in follow-up to George's appropriate comment about how do we subdivide, and we could be looking at mild disease separate from moderate disease, separate from severe disease.

I've always argued that if you do a registrational or a scientific trial in a given setting, let's say a moderate disease, the label ought to reflect what it is that you studied. If you didn't study mild disease, if you didn't study severe disease, you're extrapolating without scientific clinical data to allow that extrapolation.

So therein lies a difficulty of subdividing in too fine a way. If we subdivide out the milds from the moderates, then we're requiring separate studies in those settings in order to be able to label the product in those two settings.

It seemed to me that it was logical to separate out severe from mild to moderate, because in the severe setting we're really looking at plausibility or already established effects on endpoints of irreversible morbidity or mortality, and so it makes sense to study in that context where the mild to moderate have been pooled to enable for a more practical approach without having to look at a specific sub-trial in each of those two settings.

One quick thought. The speakers should all be at the table. John Powers should be at the table, as should other speakers from the session today.

DR. TALBOT: Yes, if I could just

comment on that, the mild, moderate, severe is really important, not only scientifically but operationally. Severe, when I think about that, is CAP patient requiring ICU care.

If you look at clinical trials recently, there are very few of those patients included, so if you separate out severe, and it's the patient needing ventilation in ICU, I actually don't know whether such a study will be done. If you then lump mild and moderate versus severe, you're lumping at least PORT 1 and maybe PORT 2 with PORT 3 and 4, but we already are thinking maybe PORT 1, anyway, shouldn't be an NI approach.

So to me the moderate group a priori is of interest for a potential non-inferiority design, because we can agree that most likely I think that there's a treatment effect, but if you're lumping them for labeling and scientific thought processes with mild, it just becomes confusing.

So, I mean, I would like to come up

with, actually, a tripartite grouping here, partly for logistical considerations, because I think it's going to be very difficult to do a study just in severe alone, and lumping severe with moderate, if by moderate you mean PORT 3 and 4 plus whatever, is also a problem, because the PORT 5s or the severe ICU patients are again a very small part of the population, and further to John Powers' point, you won't really be able to tell anything about them, anyway, so you might as well study PORT 3 and 4 as moderate, for example, or PORT 2, 3, and 4 as moderate. So, I mean, I think it's inextricably linked science and what's going to be feasible from a clinical trial perspective.

DR. SPELLBERG: Ed, can I just go back to Barry's question? I just want to make sure that I understood your answer.

Am I correct that your answer was that if you do a trial with just moderate patients, you could not assume that you would

also get a label for mild? Is that a correct way -- is that a correct interpretation or --

DR. COX: Right. I think you're asking me to answer with certainty when I think there is a lack of certainty. I think what we're here -- one of the things we're going to talk about today is the issue of, you know, what do we know about treatment effect in this milder population, and I think that's one of the issues that we're here all discussing.

So, you know, if there is a basis, if there is a reason to use that information for more, you know, severe disease as being supportive, well then, sure, that's helpful, but if you have evidence staring you in the face that, you know, starts to raise real questions about that, then I think you have to ask yourself the question of what is it, you know, are you doing.

DR. SPELLBERG: So it's a may or may not. I'm not -- I wasn't trying to pin you

down.

DR. COX: Yes, you know, that's fair.

DR. GILBERT: You were trying to pin him down, but it's okay. Okay, we've got to keep on schedule, but Rich, if you have one quick question.

DR. WUNDERINK: Just one quick comment to follow up on that as one of the token critical care people here. When we say severe community-acquired pneumonia, that means somebody who's come into the Intensive Care Unit, and the PORT score really does not reflect that completely, so you have this overlap of PORT 5s being discharged and PORT 1s being admitted to the ICU.

And I think there is a very real and practical consideration of looking at patients who are treated as an outpatient, patients hospitalized outside the ICU, and patients looking at the ICU, and I would actually make a plea that we need to study the

ICU patients, because that's where the mortality is.

That's where the significant long-term outcome issues are, and we know nothing right now. There is one study that I know of that allowed patients admitted to the ICU, and so I think that's a huge hole and is one of the really important things that may change the mortality of community-acquired pneumonia.

DR. GILBERT: We couldn't agree with you more, and the organizers had actually three scenarios. The third scenario was the severe patient in the ICU, and then when we outlined the program, there simply wasn't time to do it justice, so you're lobbying for another workshop, which will come up in a little bit.

So when you use the microphone, and I cut Rich off, please identify who you are and so forth so that on the recording of the sessions we have the speaker identified.

It's my pleasure now to introduce

Rick Nolte from South Carolina, and his presentation heralds back to the days when all of us were stuck with the patient with pharyngitis, and we didn't know if it was viral or bacterial until we got the rapid strep test, and that certainly changed clinical practice, and I think that sets the stage for Rick's presentation. Dr. Nolte.

Here's a pointer for you.

DR. NOLTE: Do I need it? Thank you.

DR. GILBERT: If you need it.

DR. NOLTE: Appreciate it. Good morning, everyone. I want to thank the conveners for inviting me. Just like we have a token critical care guy here, I'm your token clinical microbiologist. I am currently Director of Clinical Laboratories at the Medical University of South Carolina in Charleston.

What I want to do with the half-hour or so given to me today is review

molecular diagnostic approaches for detection of common bacterial and viral agents of community-acquired pneumonia, discuss in a general way the relative strengths and limitations of these approaches relative to the conventional methods, the dizzying array of methods that are used, culture, antigen detection, and serology, and then hopefully provide some evidence or at least demonstrate to you with a couple of examples of how these molecular methods may better define those subjects eligible for community-acquired pneumonia trials.

I took -- the next three slides just go through the usual suspects in community-acquired pneumonia. This is taken from the current IDSA ATS consensus CAP guidelines. Basically, as we all know, strep pneumoniae, mycoplasma, haemophilus influenza, chlamydophila, and the respiratory viruses are the most causes of CAP in patients destined to be treated as outpatients.

As you move into the inpatient but the non-ICU setting, many of the same players come up, but then we add things like Legionella and aspiration pneumonia to the list of common etiologies, and then finally, as our critical care colleague wanted to talk about, those patients with community-acquired pneumonia that move to the ICU. Again, the cast of characters changes a little bit. We're adding gram-negative enteric bacilli probably to that list.

What I'm not going to talk about today are the vast array of testing strategies that have been devised for looking for those more uncommon causes of community-acquired pneumonia, but molecular methods do figure prominently, I think, with those agents, as well. That's just in the interest of time.

Basically, we've already, I think, covered this to some extent, the specific etiologic diagnosis. In most patients, the causative agent is unknown. It's been

estimated as much as 98 percent of outpatients and in 50 percent of inpatients. Even in studies where every effort is made to determine the etiology, the success rate is often at about 50 percent.

Why is that? I think a lot of it has to do with the limitations of the sort of traditional approaches to diagnosis, and then some of it is probably due to unrecognized or underappreciated pathogens that perhaps in the underappreciated category weren't sought.

So what is some of the promise, I guess, in terms of molecular diagnostics? And by molecular diagnostics I'm really talking about nucleic acid-based diagnostics, especially nucleic acid amplification methods. They do offer the promise of increased sensitivity and more rapid results than the traditional approaches for most pathogens.

This is certainly true for the respiratory viruses. It's certainly true for Legionella and hemophilia, and also they offer

the opportunity to provide a clue to the etiology even in those patients that had prior exposure to antibiotics. These methods are currently the best alternative for pathogens that are difficult or impractical to culture like mycoplasma pneumoniae and C. pneumoniae.

When you start moving to the common bacterial agents, it may be that quantitative methods rather than qualitative detection are required to do the best job of separating those patients who may be colonized from those patients who are infected, and certainly with streptococcus pneumoniae, haemophilus influenzae, and gram-negative bacilli, those would be important concerns.

So we're talking about a combination of qualitative methods for those pathogens for which there is no normal carrier state and then perhaps quantitative methods to better define those patients that have infections with organisms that can also be colonizers.

What are some of the concerns, at least from my perspective? There is a reasonably large number of agents that would have to be sought to cover the waterfront, and, considering this, parallel testing is probably going to be impractical, so basically where I think the field is moving is in terms of multiplex analysis is a key in terms of enhancing diagnostic yield, and there are a number of approaches from a technical standpoint to that problem.

Multiplex PCR using either conventional or real-time methods, you can probably get as many as two to ten targets. One of the really sort of exciting approaches to this multiplex analysis is the so-called liquid micro arrays. This is a Luminex platform. We'll talk a little bit about that.

You can probably do up to 50, maybe as many as 80 different targets in a single PCR reaction. I mean, this is a remarkable advance, and this is technology that is

already in clinical labs, and I'll talk about one application that is FDA cleared on this platform.

And then the sort of Holy Grail, I guess, in terms of multiplexing would be the sort of solid micro arrays. There are a number of papers, research publications, that talk about using random, prime, or PCR in extensive oligonucleotide arrays to really categorize. You can cast your net as wide as you want, and you could envision a chip in the future that would cover all known respiratory pathogens.

I want to talk about two examples, first starting with this paper published by Morozumi in the *Journal of Clinical Microbiology* in 2006, where they essentially developed six real-time PCR assays with molecular beacon probes for some of the usual suspects in terms of bacterial causes of community-acquired pneumonia, and they threw in streptococcus pyogenes for reasons that

were not clear to me.

But anyway, the assay analysis time was only two hours. They compared it to serology, sort of the conventional approach for mycoplasma and chlamydia, and cultures for the other agents. What they demonstrated was a high sensitivity and specificity relative to the comparators for all organisms that they tested.

This is a real-time PCR method, and the beauty of real-time PCR is that it's inherently quantitative, that you really get quantitation without any real extra effort in that the cycle threshold in a real-time PCR reaction is inversely proportional to the starting number of target molecules, and what this graph shows is that there is a fairly good correlation between the semi-quantitative culture results and the cycle threshold in the PCR reaction, and this happens to be for streptococcus pneumoniae.

And also what's interesting in --

my animation apparently isn't working, but basically -- I'll do it the old fashioned way if I can figure out how to turn on the pointer.

In this group here, the culture negative/PCR positive group, all of these -- I think there are seven or eight patients here. All of these patients had prior antibiotic exposure, so basically it extends the ability to detect the pathogen even in those patients that were culture negative and probably culture negative because of the prior antibiotic exposure.

Oh, there it is. Here we go. Same thing for haemophilus influenzae, a very good correlation between the cycle threshold and the semi-quantitation of culture results, and again, those patients with prior antibiotic exposure were all of the patients that had positive PCRs and negative culture results.

Okay, let's move on to respiratory virus detection, because I think this is

really where the molecular diagnostic approach has significantly increased our diagnostic capabilities, just basically a little bit of review.

You know, what are the approaches that have been used? Serology, obviously we understand the problems with that. It's a retrospective diagnosis.

Rapid antigen detection, there are a variety of approaches that have been taken, but for most viruses, they have poor sensitivity and some problems in specificity, as well, with the exception, perhaps, of respiratory syncytial virus.

The culture approach, conventional cultures are too slow to have any real clinical impact. There have been tremendous advances in terms of quick or rapid cultures, so-called shell vial techniques, but we realize that some important respiratory viruses do not grow in cultures, so that's not really an option.

Nucleic acid amplification tests really have emerged as the new gold standard in this area. They provide rapid results and excellent sensitivity, and there are a variety of approaches that have been taken from single target assays to multiplex assays with two to seven targets per reaction to massively multiplexed analysis, which includes ten to 20 viral targets.

One system that we've had some experience with is made by a company called EraGen. This is essentially a three-hour process that detects 17 different respiratory viruses in a single sample. It employs some proprietary technology by EraGen and also uses as common platform the Luminex Xmap platform for the readout. There are no washes or transfer, and this really can be a high throughput system.

Basically, the technology is laid out on this slide here. It's a reverse transcription step, because most of the

viruses that we're seeking are RNA viruses followed by PCR, and then a step called target-specific extension, where more target is made along with this capture probe that is attached to the primer that's used in this step, and then there is a complimentary sequence on these polystyrene beads that captures the specific PCR product.

The key to the detection is this Luminex Xmap technology, which involves a series in this case of 100 color-coded beads. Each one is individually addressable, and on those beads you can link specific oligoneucleotide capture probes, and the whole process is read by a dual laser flow cytometer, one that identifies the specific bead, the other that reads any signal that might be associated with it.

Basically, what the panel looks like in this iteration, there were 17 different viruses, human metapneumovirus, influenza A and B, parainfluenza virus 1, 2,

3, and 4, respiratory syncytial virus types A and B, the respiratory adenoviruses belonging to the groups B, C, and E, human rhinoviruses, three coronaviruses, OC43, NL63, and 229E, and the appropriate sort of internal positive controls.

As I mentioned, this is a high throughput system. This shows the output for the influenza A virus assay. I think there were something like 180 samples examined in this run. What we have here are the results for the influenza A portion of the assay. You can see they are all well separated from the threshold values down here. These are all the negative samples.

These are the internal positive controls, if you will, an RNA-positive control and a DNA-positive control, and for the most part all of these are successful. There are a few RNA-positive control failures here, so the test has the right kind of controls to give you confidence in negative results.

Basically, so what we went through, essentially a methods comparison where we looked at 354 specimens primarily -- all from adult patients, primarily from hospitalized patients, and many of these specimens were lower respiratory tract specimens.

Many of them were BALs, and these are looking at the viruses that were in common between the culture-based method, the R mix method, and the molecular method, the PLx method, and you can see with the exception of influenza A there were comparable yields for the other viruses on this panel, but you see the dramatic increase in the number of positive samples for influenza A.

Moving on to the viruses that aren't normally sought by culture and that we didn't have a culture backup for, you can see that there is significant -- we found nine patients with human metapneumovirus, 15 with rhino virus, and three patients with these coronaviruses, NL63 and OC43. So looking at

the overall increase in diagnostic yield, it sent from 22 percent by culture-based methods up to 33 percent with the molecular method.

Also, an approach like this also gives you the opportunity to identify those patients with mixed viral infections, and in this particular situation we had two patients, one with a coronavirus and a rhinovirus and the other with the human metapneumovirus and a rhinovirus.

There are several different manufacturers who are approaching the problem on this Luminex platform. TmBiosciences is another company that is partnered with Luminex in producing these kinds of panels. This shows you their test menu, and basically this is technology that was first described by Jim Mahoney in the *Journal of Clinical Microbiology* article listed here.

The good news here is this is the -- this assay the first of the year was cleared by the FDA. It's the first example of a test

with this kind of power to be cleared by the FDA, and that's pretty good news, I think.

There is another approach, a very similar approach, based again on this Luminex platform that incorporates bacterial targets, as well as viral targets, and it's this kind of approach that is intriguing, I think, in terms of really expanding the diagnostic capability.

This company, Genaco, is partnered with Qiagen, and they have two panels, if you will, one and two. The first panel goes after DNA targets in terms of the PCR reaction and covers the usual suspects in terms of community-acquired pneumonia, the bacterial pathogens, and adds adenovirus in there, because it's a DNA virus rather than an RNA virus, and then the second panel covers the usual suspects, if you will, in terms of respiratory viruses, the RNA viruses.

Part of the problem is that there are precious few FDA-cleared diagnostics,

molecular diagnostics for respiratory pathogens. Certainly Mycobacterium tuberculosis is covered.

Recently there has been a molecular-based assay for Legionella pneumophila approved by the FDA. The respiratory virus panel. The Luminex panel produced by TmBiosciences recently received FDA clearance, as I mentioned, and also there is a real-time PCR produced by a company called Prodessa that covers the three viruses listed here, influenza A, influenza B, and RSV.

From this brief overview, what can we conclude? Molecular diagnostics, I think, do have the potential to better define subjects eligible for these kinds of trials by improving the diagnostic yield and decreasing the time required to identify etiologic agents.

This analysis can be completed in a matter of hours rather than a matter of days,

particularly if you're focused on antibacterial agents and quickly identifying those patients with viral infections that might not be appropriate for your clinical trial.

The lack of FDA-cleared diagnostics for the common bacterial pathogens is a serious limitation. This presents problems in terms of whatever assays might be developed for use in such trials, would sort of lack the standardization. Also there is this issue of availability without FDA-cleared diagnostics.

So one of the things that I think is important as we think about new trials for community-acquired pneumonia is perhaps there should be consideration given to the development of companion diagnostics along with the drugs, because right now you can't go to the shelf and pull off a set of reagents that are going to do what might be required to get the biggest bang from your buck.

Also, the problem -- it's not on

here, but part of the problem is transitioning from the sort of culture-based methods to molecular diagnostics methods. You're still probably going to have to capture an organism to determine its in vitro susceptibility.

You can approach that from a molecular standpoint, but that gets hopelessly complicated in terms of the number of -- well, in terms of the -- our lack of understanding in many cases of the genetics of bacterial resistance, particularly with new agents and, you know, covering the waterfront in terms of all of the agents that you might consider as an etiology and making sure that you completely cover all of the possible antibiotic resistance mechanisms to those drugs. So I think that brings me to the end.

DR. GILBERT: Thank you very much, Rick.

DR. NOLTE: Thank you.

DR. GILBERT: I suspect there will be many questions and comments. I'll just

start out. So I know someone is going to ask you about colonization versus invasive disease. So you showed some quantitative results with the real-time that related to prior antibiotics, et cetera, but quantitation might also help us, would it not, with respect to colonization versus invasive disease?

DR. NOLTE: Absolutely, and I think that's -- I know Dr. Klugman is going to talk specifically a little more about the pneumococcus in the quantitative PCR story, but basically it's, you know, it's like quantitative cultures. It's going to help you to some extent define the bacterial burden and perhaps give you another marker in terms of treatment response and watching the quantities of those organisms decline early in the treatment process.

DR. GILBERT: And then the Luminex platform, I guess that's the one that's FDA-approved, at least for the 12 pathogens. What does it take to get approved? I mean, what

were the patient populations that were studied, or was this just all in vitro studies that led to the approval?

DR. NOLTE: Yes, they're methods comparisons and comparisons to conventional technology that was in use, basically culture-based methods. It wasn't -- it's -- you know, diagnostics and drugs take different paths through the FDA, and that's one of the things that I think there's an opportunity here to do some good, because there's an awful lot of talk about companion diagnostics for other types of drugs, for cancer drugs, all the talk about pharmacogenomics and characterizing people's cytochrome P450 genotype as a drug goes through the FDA and having those tests migrate with the drug through the FDA, and I think here is another opportunity, because clinical laboratories really don't provide --

I mean, the technology that's available is really we all recognize the limitations of it. There is an opportunity

here with some of the newer technology to migrate these diagnostics along with the trial of new drugs.

DR. GILBERT: Just to be clear, we've got two other questions. None of these tests have been vetted clinically. I mean there's no correlation that's part of the approval process with the clinical disease.

DR. NOLTE: Typically what happens in the trial of a diagnostics like this is samples are submitted to the clinical laboratory for, you know, for whatever reason, suspected, you know, in this case suspected viral, you know, respiratory viral infection, and then there is patient information collected, but there is no -- the test is approved for detection.

DR. GILBERT: Got it.

DR. NOLTE: Okay.

DR. GILBERT: Okay, I think Dr. Psaty was next.

DR. PSATY: Yes, I think the

diagnostics provide an interesting opportunity here. Let's say we designed a clinical trial that included the diagnostic as an eligibility criterion. The proper generalization of that trial to clinical practice would be then to apply that same diagnostic and use the antibiotic as it was used in the trial, and so are you envisioning at all that the indications for the drugs will include potentially the use of the diagnostic test?

DR. GILBERT: Well, Ed?

DR. NOLTE: Is that a question for me or the panel?

DR. GILBERT: I think that's a question for the Agency.

DR. COX: Yes, there are instances where, you know, use of the test is integral.

DR. GILBERT: Can we get Ed's mic to work here?

DR. COX: Sorry. I'll pull it a little closer. There are instances that arise that, you know, where the use of the test is

integral to understanding the risk and benefit of the product.

You know, there may also be other instances where, you know, you're using the diagnostic for the purposes of, you know, enriching the clinical trial, but, you know, you may not necessarily need to have the diagnostic in order to affect the risk benefit.

I mean, there may be other reasons why, you know, other scientific reasons that you'd be willing to generalize beyond just specifically that population where, you know, that had the test in the clinical trial, but you raise a good point, and, you know, how these tests can be used, you know, best and really the feasibility of their use for, you know, design in clinical trials.

DR. TEMPLE: Yes, well, it depends a little bit on how desperately you don't want to give it to somebody who doesn't need it. There are some examples in oncology.

For example, the drug Herceptin is recommended only for people who have the appropriate receptor on their breast tumor, and one of the things that drove that determination was how toxic Herceptin is. It causes heart failure. There are other drugs that could be similarly directed where we've not been as insistent, and I think that's part of it.

I just want to make one observation, which is that while trying to identify the people with the relevant disease as early as possible so you don't expose them to a drug that can't benefit them is good, from the point of view of interpreting the trial, dropping them out later is okay, too.

It's a baseline characteristic. Antibiotic trials always drop people who don't have the right organism. I mean, it's been done for 50 years that way or 20 years, 30 years, so that would be an interpretable trial, too.

It's a baseline characteristic. You could if you diagnosed them appropriately and blindly and everything like that, drop them out later and make the end point the people who do have a bacterial disease.

DR. GILBERT: It would impact sample size rather dramatically.

DR. TEMPLE: Absolutely. I mean, you'd still -- it would be better to exclude them. I'm just saying if you can't get people to do it, you can still have an interpretable trial.

DR. FLEMING: But you're just to follow up --

DR. TEMPLE: Including -- but including them without finding out what they have is just a way to always win in a non-inferiority study, so you do have to find out who has the -- who doesn't have the right disease some time.

DR. FLEMING: But, Bob, you've got to follow up on what you had said, and that is

if you drop them out, you've got to be confident that the people that are the unintended people are in fact not getting harm, and in severe sepsis trials for years --

Let's say if you're looking at gram-negative sepsis, and you in fact then want to do a subsequent analysis on those that are found gram-negative, the complimentary group you in fact need to be ensured you're not providing any harm to, because they in fact have received the intervention at all.

DR. TEMPLE: Absolutely, and the safety analysis includes everybody randomized, not just the ones you decided to study effectiveness in, but you could have your primary end point being the people who have the disease. In fact, in a non-inferiority trial that's absolutely crucial. Otherwise, you'll always win.

DR. GILBERT: Dr. O'Neill.

DR. O'NEILL: I just wanted to put a plug in for the CDRH folks who are likely not

here. There is a guidance on the co-development of diagnostics in drugs in the works, and it's probably been in the works for two years.

It's sort of -- in a sort of --on hold, because there are some differences of opinion on the level of standards that you need for a diagnostic, and it was interesting to hear that one is approved. It would be nice just to have a discussion of the evidence behind why that was approved and why the other guys aren't approved.

I think there is a history of convenience sample testing, which doesn't rise to the level of evidence here. A lot of the problems that the CDRH folks have is they want legitimate studies, not unlike we want control clinical trials that allow you to establish the sensitivity and the specificity and the positive predictive value of these particular tests in different populations where the prevalence is different or whether the

phenotypes are different in terms of the way you come.

So they need that described well, so anybody who is interested in this game who wants to package sort of diagnostic along with the drug, he needs to be thinking holistically in terms of how do you design and get two-for-one. How do you design your clinical trials, and how do you get the other information on the diagnostic?

People aren't thinking that way right now, and this discussion is going on in the biomarker world, not in this world, but the biomarker world, with saying "Come up with some magic biomarker, and enrich -- and if you're lucky enough to enrich your clinical trial population, and then you go and put your money on that biomarker positive subgroup if that -- if treatment effect is demonstrated there, well, you're home free."

The issue is how do you operationally implement that if you go out and

sort of allow people to be entered on the basis of that? I've been thinking about this in a number of other ways.

If you look at the recent publicity on the Gail model for the use of tamoxifen, the Gail model was something that's actually in the label of tamoxifen, and it was actually used for the entrance criteria into tamoxifen, and it essentially was a -- it's a logistic regression model that plugs in three or four co-variates that a woman has and sort of says if your probability of breast cancer in five years is less than, let's say, eight percent, you're eligible or you're not, so there's sort of a yes or no kind of a thing.

And that model can be viewed as a diagnostic test in that sense, and there was evidence that it really underestimated substantially the probability for black women, and so they're reratcheting that, because there's a lot of implications for getting that wrong.

The reason why I'm telling you this story is because CDRH worries about that, and they're worried about the sensitivity and specificity and the positive predictive value, so you're going to have to marry both of these ideas if you want to get some home runs here, and I don't think that conversation has really gone on enough. I don't think there is enough clinical trial drug development strategies that marry both the design for the diagnostic as well as the design for the trial.

DR. NOLTE: There are a couple. I mean, I actually serve on CDRH advisory panel, the microbiology devices panel, and, you know, I go to meetings, and I hear a lot of talk about companion diagnostics.

I don't hear about it so much in the infectious disease arena, and, I mean, in addition to the problems that you have with FDA, parts of FDA working together, I mean, most of the companies are in the diagnostic business. They're not in the drug business,

so you're now talking about, you know, partnering, making alliances between different companies that may not have the same interests, but I think it's a marvelous opportunity.

I mean, there are a couple of examples of it in the infectious disease world, and I think, you know, the development of antiretrovirals and the co-development of viral load tests for HIV by Roche and for HCV, as well, those -- a lot of information was gained about the diagnostic from the clinical trial of the antivirals, and I think, you know, you can, as you said, you can mine that same data set and get information that you need for the drug trial, as well as the diagnostic trial, if you put it together right.

DR. GILBERT: Thank you. So, now we want to address one example of a surrogate marker, if you will, and prospects for calcitonin as a new biomarker. We've asked

Michael Niederman, who is Chairman of the Department of Medicine at State University of New York, to help us review where we stand there, and I hope, if I --

DR. NIEDERMAN: If you click on that, it'll go on.

DR. GILBERT: Yes, but I've got to get my clicker in the right place. Thank you, Michael.

DR. NIEDERMAN: Well, thank you. It's been a very interesting discussion, and I thank all of the organizers for giving me an opportunity to be here today.

I don't claim to be an expert in biomarkers, and I've had a good time looking at this literature, and I'll try to synthesize for you what I've learned in reading this literature and how this material might be incorporated into the thinking about designing a CAP trial.

My major interest is in CAP itself and in the approach to diagnosis and therapy,

but I do think that the role that biomarkers could play in trial design is very important, and, as I said, at the end I'll try to give you a synthesis of where I see this literature potentially helping us.

So why should we even think about biological markers? We heard from Dr. Powers about we have clinical endpoints. Why not simply use them for entry and for evaluating the outcome in clinical trials? And clearly there are many issues with the use of clinical parameters.

Clinical features depend on the host response to infection. I don't think we focused on that tremendously today, but the host response can vary by organism. It can vary in relation to prior therapy and certainly can vary in relation to the host, and I don't think that we've really considered that in a lot of our discussion today.

The issue of age came up, but comorbidity, genetic polymorphism, and the

immune response all can lead to variable clinical presentations, which we then translate into severity measurements. The severity of the illness itself can certainly affect the clinical presentations.

Certainly, as we heard in the discussion between bacterial and viral infections, clinical information isn't specific for infection in general or for specific etiologies, and many of the clinical features that we relied on, for example, the chest radiograph, may give us information that's much too late in the course of the disease to truly identify all the patients. We've all dealt with patients who have initially negative chest radiographs who indeed have community-acquired pneumonia.

A number of biologic markers have been looked at for pneumonia. The pro-inflammatory cytokines have been primarily research biomarkers, TNF alpha, IL-1, and IL-6, but I think it's important to understand

when we talk about procalcitonin that these pro-inflammatory cytokines can actually stimulate acute phase reactants like procalcitonin. Anti-inflammatory cytokines, again, probably not very useful for the purposes that we're talking about today.

The acute phase reactants that have been studied extensively are C-reactive protein and procalcitonin, which I'll talk about in a moment. S-TREM, which is a member of the immunoglobulin superfamily, has had a little bit of study.

It probably is not practical, either, as I'll show you in a moment, for the purposes that we're looking at today, and there are a variety of other biomarkers that have been looked at that I won't focus on in the discussion.

The interest in S-TREM I think became widely known after this *New England Journal* study several years ago. TREM is the triggering receptor expressed on myeloid

cells, which is up regulated by the presence of bacteria and fungi, and this immunoglobulin is shed by the membranes of activated phagocytes in the soluble form, then is present in body fluids.

The study in the New England Journal measured S-TREM in bronchoalveolar lavage fluid of patients suspected of having pneumonia. They included patients with more severe CAP, then a later associated pneumonia or no pneumonia, and their validation of the presence of pneumonia was to look at a clinical pulmonary infection score and quantitative cultures of bronchoalveolar lavage.

In the major finding is shown here, then in the BAL fluid, patients with CAP, patients with VAP had much higher S-TREM levels in bronchoalveolar lavage than patients without pneumonia. This is a potentially rapid test. It did correlate with a bacteriologic gold standard diagnosis.

Problem for the purposes that we're talking about today is that this was only validated in mechanically ventilated patients, and it required a bronchoscopy with a bronchoalveolar lavage to get a sample, and in more severe pneumonia it was clear that it was valuable. It's less clear to me from reading this study if we could use this in milder patients and earlier forms of pneumonia.

C-reactive protein has been studied in a number of settings. I'm only going to highlight a couple here. This was an emergency department-based study, 168 patients with acute cough. Twenty ultimately had pneumonia. The others had diagnoses of bronchitis, asthma, and upper respiratory illness, and with a cutoff number that they chose, they had a 70 percent sensitivity and a 90 percent specificity for CAP.

They did not relate it to illness severity, and interestingly, this study did find that if you combine a biomarker with

clinical parameters, you could improve some of the diagnostic ability.

Another study of C-reactive protein looked at 200 patients with CAP, and in general the patients with CAP had a substantially higher CRP than healthy controls or patients with respiratory tract infections that were not CAP. There was some relationship to etiology in that higher levels were seen with pneumococcus and Legionella and lower levels with viruses and atypicals, and I think that this could be potentially very important in trial design.

We certainly would like to identify patients who can potentially benefit from antibiotics versus those who cannot, and in general there was also a correlation with severity of illness, comorbidity, and need for admission. So this was a promising biomarker, but as you'll see in some of the more recent studies, this is probably not quite as valuable as some of the data in procalcitonin.

When I first heard about procalcitonin, it didn't make a lot of sense to me as to why this was a biomarker of infection, because in medical school, when I had learned about procalcitonin, I learned about the constitutive production by the thyroid, but I think if you're interested in this topic, there's a very good current review that explains some of the science behind procalcitonin pointing out that there are three potential sources of procalcitonin in the body, the constitutive production by the thyroid, the parenchymal tissue, particularly in the liver, whereas this can be an acute phase reactant that comes on relatively quickly and stays for as long as a week after stimulation in the setting of acute infection, and then a more situational production by monocytes.

Procalcitonin has been referred to as a hormokine, meaning it can be either hormonally expressed in the neuroendocrine

cells or a cytokine-like release in response to microbial toxins or a host response, and I think one of the appeals of procalcitonin is that levels can be increased dramatically.

They can be produced by the liver primarily in this parenchymal form. It can be produced by monocytes but not nearly to the extent that it's produced by parenchymal cells.

Levels can rise 100,000 fold above normal in the setting of infection, and the stimulus for procalcitonin release can be microbial toxins, which, again, makes it somewhat specific for bacterial infection, but also importantly the host response itself stimulates PCT release, and the viral host response actually inhibits PCT release.

So for the kinds of issues that we're looking at today, at least theoretically PCT could be very important, because it does respond to the presence of bacteria. It responds to the inflammatory response itself,

and it's down-regulated in the presence of viral infection.

In reading about this, it's clear that there are a variety of different assays, and if we were going to try to apply this to clinical trials, we have to pay attention to which assay we're actually using. The Kryptor assay is the one that in the literature appears to be the most accurate.

Having said that, very few people have studied the Kryptor assay, and this group from Switzerland, which has done most of the work, has done their work with the Kryptor assay. It has not been in the hands of many others other than these individuals, and so when I tell you the conclusion that they have that the Kryptor assay is better than the other approach, the LUMI test, for example, that's based on the fact that they've used this assay in their studies and have gotten better results than other people who have used other assays in their studies.

It's a sheep-based polyclonal anticalcitonin antibody. It can detect levels as low as three-fold greater than normal, and it's got a very low sensitivity in terms of detecting very low levels of procalcitonin.

Results can come in an hour or less, and the group from Switzerland has studied the use of procalcitonin using the Kryptor assay, and what's been most impressive about their studies is that they've actually not correlated it just with microbiologic data like we saw, for example, with the PCR type testing, but really they've correlated it with clinical management, and they've used it for the purpose of antimicrobial stewardship in patients with suspected respiratory tract infections.

They've used it in community-acquired pneumonia to reduce the number of patients who actually have radiographic infiltrates and not give antibiotics to certain ones who either have viral infection

or non-infectious causes of their lung infiltrates.

The issue of duration was raised. PCT has been used to determine the duration of therapy in community-acquired pneumonia, and it may also be a prognostic as well as diagnostic marker, but it's unclear if it's quite as good prognostically as diagnostically.

This is the type of approach that the Swiss have used when they've set up their studies. They've defined four different levels, and we were talking about mild, moderate, severe, et cetera.

This is a different approach using a biomarker less than .1 micrograms, .1 to .25, .25 to .5, and greater than .5, and if they've received no antibiotics, or, I'm sorry, if the PCT, rather, is in this low range, they don't recommend giving antibiotics. If it's in the high range, they do recommend giving antibiotics.

And I think, quite to their credit, when they've designed these studies they've had opt-outs for the clinicians, so if you have a very low PCT, and the recommendation would be to not give antibiotics, if there was respiratory or hemodynamic instability or if the score was less than .1, if the patient fell into very high PSI or CURB scores where if it was .1 to .25, again, they give criteria for overriding the order not to give antibiotics.

On the other hand, when the levels are very high, they recommend reevaluating the PCT after several days of therapy, day three, five, and seven, and with these cutoffs, those then become the cutoffs for stopping antibiotics. So if you started at, say, above .5, but you feel down to .1 on day three, they would recommend stopping antibiotics.

For outpatients, they recommend using PCT levels serially to guide duration of therapy for up to seven days but certainly

often less than seven days, and if the levels start very high, rather than using these as the cutoffs, they recommend to use only 80 -- dropping to 80 to 90 percent of the peak value would be enough to allow discontinuation of antibiotics.

In the review article, they highlighted three studies. I don't know if you can read that, but I'm going to go through two of the relevant studies for pneumonia. They had very nice acronyms for them.

The two pneumonia studies are the ProRESP study, which is an ED-based study that included a variety of patients with pneumonia and other diagnoses, and in the setting of upper respiratory illness mixed with pneumonia, all presenters with respiratory infection to an emergency department. The use of PCT led to a 44 percent reduction in the use of antibiotics.

ProCAP study included patients only with radiographic community-acquired

pneumonia, and even in that setting, they withheld antibiotics in 14 percent of the patients using a procalcitonin guidance, and they substantially reduced the duration of therapy from 13 days down to six days using procalcitonin guidance.

The middle study I'm not going to talk about. It was a COPD study and an exacerbation of COPD. They also used procalcitonin to guide the decision-making about use of antibiotics.

The ProRESP study was a prospective cluster randomized single blinded, meaning that the investigators did know the results of the procalcitonin. In 243 patients with lower respiratory tract illness, half got standard therapy, half had the procalcitonin guidance based on the parameters that I showed you, and what you can see here, if you look at the standard group in gray, the procalcitonin group in the darker bars, there was a slight reduction in the usage of antibiotics in

patients with established community-acquired pneumonia.

The majority of the benefit occurred in patients with bronchitis and other respiratory diagnoses. So a lot of the withholding of antibiotics that occurred in this study didn't really occur in community-acquired pneumonia, although there was some withholding of antibiotics in CAP patients.

Again, the PCT group got 44 percent less antibiotics than the control group. They had a shorter duration of therapy. This was not as dramatic as in the CAP study.

It was 10.9 versus 12.8 days, and they only could find one bacteriologically positive community-acquired pneumonia patient who by PCT guidance was not given antibiotics, and still, in the absence of antibiotics with positive bacteriology and positive radiology, that patient recovered without antibiotic therapy.

When patients had community-

acquired pneumonia as a reference range, the PCT on average was between 3.9 and 4.6, and oftentimes it's much higher than that, but again, consider that number in relation to the cutoffs for the decision of using antibiotics.

The ProCAP study was, I think, a much more impressive study. It was 300 patients with radiographic community-acquired pneumonia, again randomized to PCT-guided decision-making about antibiotics versus standard care, again, the same algorithm and the same cutoffs for whether or not antibiotics get used.

They measured PCT on admission. If they weren't sure about the withholding of antibiotics, they could withhold and then repeat six to 24 hours later so that there was a chance to treat in situations of uncertainty, and then they repeated it on days four, six, and eight.

Twenty-eight percent of the PCT study -- twenty-eight percent, rather, of the

population had a PCT value less than that 0.25 cutoff. In other words, if you followed the algorithm strictly, 28 percent should not have received antibiotics.

In the end, 15 percent with radiographic pneumonia in the PCT group had antibiotics withheld, and I think that that's still a very impressive number, given that most of these patients were admitted to the hospital. Ninety-seven percent were actually in the hospital with radiographic pneumonia.

They reduced the duration of therapy from 12 to five days. That's what's illustrated in the graphic here, and, again, I think that this shows you the potential to use this in clinical practice.

Subsequently, they combined the ProRESP and the ProCAP study to look exclusively at the patients who had community-acquired pneumonia in both studies, and that represented a group of 373 patients who had radiographic community-acquired pneumonia.

Within that group with radiographic community-acquired pneumonia, 20 percent -- 20 patients, rather, had non-infectious diagnoses. They had radiographic pneumonia but, in fact, had other diseases that led to the radiographic abnormalities, and 24 had no -- had other diagnoses, particularly viral infection, and what they then did is look at the ability of the procalcitonin level in combination with a highly sensitive C-reactive protein assay and clinical features including the radiograph to predict the presence of pneumonia, and they found, again, with looking at the area under these curves that the procalcitonin added with highly sensitive C-reactive protein and clinical evaluation was the most sensitive model possible for detecting radiographic and clinical pneumonia and distinguishing those patients with pneumonia from the 44 patients who had radiographic infiltrates but did not, in fact, have bacterial pneumonia.

They're in the process now of designing a large multi-center trial, which is outlined here. This trial has not been done. The reference is here for those who are interested, but they have designed a prospective randomized open intervention in over 1,000 patients in six Swiss hospitals, and they're going to use again the same approach of PCT guidance and randomization by center and type of respiratory infection to use as their primary end point treatment failure at 30 days and again the decision-making to be guided by procalcitonin. Secondary endpoints in this proposed study are antibiotic exposure, rate of hospitalization, cost-effectiveness, and time to clinical stability.

One of the issues that we would consider is the ability to use procalcitonin to separate bacterial from atypical pathogens. It's a little bit harder to get good data on this.

This is one study that I came across of only 30 patients with community-acquired pneumonia. Ten had documented atypical pathogen infection. There is a typo in your handout. That's 20, not 30, bacterial infections, including three that were bacteremic.

They used the other PCT assay, the LUMI assay, and they found that the PCT was substantially higher for the bacterial pathogens. In this testing, 7.6 versus 0.8 for the atypical pathogens.

It becomes a little problematic with these numbers, because the mean value of 0.8 is still above that threshold of 0.5, so you might end up treating atypical pathogens, but this might be a potential way to get around the discussion we've already heard about do you have to include atypical pathogen coverage in a CAP trial. Maybe you could use cutoffs like this to address that issue.

On the other hand, in their trial

C-reactive protein and clinical parameters were not helpful in separating out bacterial from atypical pathogens from one another.

There is the potential to use C-reactive protein -- I'm sorry, procalcitonin serially to predict prognosis in community-acquired pneumonia, and although this is the population that Rich was talking about, severe pneumonia, I think it raises an important concept.

For patients admitted to the ICU with severe community-acquired pneumonia, if you looked at PCT levels on day one and then subsequently on day three, patients who survived compared to those who died started with a lower PCT level, and it dropped by day three. Patients who died started with a higher PCT level, and it rose by day three.

And so serial measurements in procalcitonin may have some value, not only for stopping therapy, but some prognostic value, and there have been a number of

studies, this just being one of them, that have correlated initial measurements of procalcitonin with the PSI score and with other outcomes in community-acquired pneumonia.

This study had 185 CAP patients who had PCT measured during the first day. Most were inpatient, a few outpatient, but you can see that patients in PSI Classes I and II had a significantly lower PCT than patients in PSI Classes III through V, and the development of complications throughout the course of their stay was also associated with a higher PCT level.

So a low PCT level, whether it indicates viral infection, other diagnoses than CAP, or even a bacterial infection, still correlates with a very low frequency of complications.

So, to conclude, I think PCT is probably the most promising biomarker to at least define the need for the use of

antibiotics in lower respiratory tract infections, including CAP. It seems to be very valuable for separating out bacterial from viral CAP, but it does appear from what I can read that the Kryptor assay is probably more valuable, and in that regard it probably needs more validation by multiple investigators. Most of these studies have not been done in the United States, haven't been done with our thought processes and our management of patients.

If you use PCT with high sensitivity C-reactive protein, in the one study it could enhance the value of clinical features for predicting radiographic CAP, and it can also identify patients with the worst prognosis in CAP. Higher values correlate with a higher PSI score, and serial measurements, as I showed you, may have some prognostic value.

So the last slide is a speculation about how we might use this information in CAP

trials, and we might, for example, if we were going to be measuring this PCT regularly, omit anybody with radiographic CAP and a PCT value that's low. That could be less than .1 or possibly less than .25, provided they don't have any of those other criteria for which we'd want to treat them in a trial, human dynamic instability, desaturation, or higher prognostic scoring groups.

By omitting these patients, again, you address an important issue that's already come up today. You take out of your antibiotic trial patients who get no benefit from antibiotic therapy.

If you were hellbent on doing a placebo-controlled trial, I guess you could do a placebo-controlled trial safely in this population, probably to prove that antibiotics aren't necessary, but I don't think -- I know clinically what I would do with that type of data.

In the outpatient with CAP, if you

wanted to design a superiority trial, then potentially you would take people with the highest PCT values, pick that group of .5 or higher, since at least that's the group that has the greatest risk of a poor outcome, and you might see discriminating value if you were trying to design a superiority trial.

For a non-inferiority trial, PCT level of .25 or greater could be an entry criteria, because at least you could be sure these are patients who might benefit from antibiotics, but by including patients with all degrees of likelihood of complications, to me this population would be more appropriate for a non-inferiority trial, and again, if you're looking in clinical trials, serial measurements of PCT that drop could be another surrogate marker, particularly in a superiority trial.

If you wanted to, again, discern a difference between two different therapies, potentially having one therapy lead to a more

rapid drop in PCT is a legitimate surrogate marker and may be more sensitive than some of the clinical markers that we're currently using. Thank you.

DR. GILBERT: There's one approved methodology in the United States. Is that the Kryptor method, or is that the other method?

DR. NIEDERMAN: I think it's the other method. I don't think the Kryptor is approved yet.

DR. GILBERT: Dr. File?

DR. FILE: Thanks, Mike, for that very nice review, comprehensive review. As we know, some of our antimicrobial agents have immunomodulatory effects, and is there data on, for example, what effect the macrolides have in vitro, for example, on the production from either parenchymal cells or monocytes of PCT, because if we wanted to use this as a marker of response, we'd have to take that into account?

DR. NIEDERMAN: Yes, I don't know

that data. There may be, and I don't know. If there's somebody else who knows that data, please feel free to chime in, but I did not see that, other than the relationship of inflammatory.

In other words, I could imagine what is known is the effect of other inflammatory mediators to up and down regulate, so to the extent that an antibiotic would initiate a specific immune response, that could indirectly affect it, but I don't know the specifics of an antibiotic correlation, specific antibiotic rather than the general effect of therapy.

DR. GILBERT: Yes, Dr. Psaty?

DR. PSATY: Very nice presentation. I really appreciate that. In your last slide, you indicated that -- thank you -- that there is no benefit of antibiotic therapy in this group. How solid is that information, or is that a legitimate area for study with placebo-controlled trials?

DR. NIEDERMAN: Well, you've seen the data, and you can decide how solid it is. I think, having read these studies, I would be very comfortable not using an antibiotic in a patient who had a PCT level less than .1 with none of these findings, even if they had a radiographic infiltrate.

I would have no trouble doing a placebo-controlled trial in that study, in that setting, but I'm not sure what I'd be proving, because I think that we have pretty good likelihood that those patients either have a viral infection or don't have bacterial infection that's going to lead to complications. I think it's a little hazier between the .1 and the .25, but I do think that those are the data that exist right now.

DR. GILBERT: Okay, I have to take the Chairman's prerogative here. I know there's lots of other questions, and we'll ask Michael to stay at the podium to answer them, but I think we need our full physiologic

break, so we'll start promptly again at 10:35. Thank you all.

(Whereupon, the foregoing matter went off the record at 10:24 a.m. and resumed at 10:37 a.m.)

DR. FLEMING: I think it's time to reconvene. If we could all come back, grab our seats, we're very pleased to have Michael Fine to talk to us about how severe pneumonia is assessed through the PORT scores, and Michael is coming to us here from the Pittsburgh Health Care System.

DR. FINE: Again, thank you very much for the opportunity to talk at this symposium today. The title for my talk is "The Pneumonia Severity Index: A Decade After Development."

So I'd like to try to address several questions in the next 15 minutes. First, what is the Pneumonia Severity Index? And we'll call that the PSI. Ten years ago, what motivated the development of the PSI, and

do those motivations still exist today? And I believe they do.

How was the PSI derived and validated? What is the effectiveness and safety of the PSI in guiding clinical practice, which was one of the major motivating factors behind its development? And what are some other applications, caveats, and limitations of the PSI?

One of the things that I will not do today due to the restriction in time is to do comparisons of the PSI to other severity adjustment models for pneumonia, though there is a literature on that topic.

So what is the PSI? It's a prediction rule for prognosis of community-acquired pneumonia, which we'll call CAP, that's based on 20 variables that are routinely available at the time of patient presentation. In addition, it's also a decision aid that stratifies patients into five risk classes, identifying a low-risk

subset that can safely be treated in the outpatient setting.

Since its derivation and validation, the PSI has been cited in over 1,300 publications, according to Google Scholar, and has emerged as the reference standard for risk stratification for CAP. Just to put that in perspective, the 2000 IDSA guidelines for pneumonia, which have also been very widely cited, have been cited in approximately 1,000 scientific publications.

So what was the original motivation for developing the PSI? Decision aids are most useful when clinical decision-making is complex, clinical stakes are high, and opportunities for cost savings exist without compromising quality of care for patients, and these were the exact circumstances for the PSI.

As you all know, pneumonia has a high mortality. It's the sixth or seventh leading cause of death combined with influenza

in the country, and it has a wide range of mortality, from less than one percent for outpatients to greater than 30 percent for ICU patients, but how do we determine which patients have a one percent mortality from those that have a 30 percent mortality when they present to physicians' offices?

In addition, Jack Wenberg and others showed that there was wide variation in admission rates for patients with similar severity of illness at the time of presentation, suggesting that if physicians did not objectively quantify risk but used a lot of subjective decision-making in making key decisions about who should come into the hospital versus who can safely be treated at home.

We also did a number of studies that showed that physicians actually overestimate the risk of death for low-risk patients with pneumonia, and such overestimation actually led to hospitalization

of low-risk patients.

And, finally, data that our group generated, as well as Mike Niederman and others, there is an extremely differential in the cost of inpatient versus outpatient care. The ratio is about 20 to one, and in 1994 dollars, the estimates were about $7,000 for the cost of a typical treatment or a typical episode of inpatient pneumonia versus $350 to treat a patient as an outpatient.

In addition, there were prior prognostic models at that time, but they had limitations, and I'm not going to go over all the limitations, but they're shown on this slide.

So the PSI was developed by the Pneumonia Patient Outcomes Research Team, which was a research team and research project that was funded by the Agency for Healthcare Policy and Research. It was one of numerous PORT studies that focused on common diseases, diseases where there was wide variation in

practice patterns, had significant morbidity, mortality, and cost of patient care.

The purpose of the development of the PSI was to develop a clinically applicable prediction rule for short-term mortality in patients with CAP. We hypothesized that low-risk patients can be identified at the time of presentation using readily available clinical information.

The derivation of the PSI was performed as part of a retrospective cohort of pneumonia patients from the 1989 MedisGroup's Comparative Hospital Database. These patients came from 73 hospitals and 23 states. The patients were 14,199 patients with a principal diagnosis of pneumonia who were adults defined as age 18 or greater.

We considered as predictor variables 20 variables that were independently associated with mortality and a prior pneumonia specific severity model that our group had developed, and our primary outcome

for the derivation was hospital mortality within 30 days.

We had the luxury of validating the PSI in two independent cohorts. The first was a second MedisGroup's cohort, this time consisting of 38,000 patients with pneumonia that were admitted to 187 Pennsylvania hospitals during 1991.

Again, this second validation cohort was restricted to inpatients, and the outcome measure was hospital-based mortality. In addition, we had the multi-center PORT Cohort Study that we could use to validate the PSI.

These patients were 2,287 that were prospectively enrolled between 1991 and 1994 from five medical centers in three cities. There are two medical centers in Pittsburgh, two in Boston, and one in Halifax, Nova Scotia.

In contrast to the MedisGroups validation, this validation cohort included

both inpatients and outpatients, and we had the ability to extend our outcomes assessment to 30-day mortality outside the hospital, as well as to assess a variety of other clinically relevant adverse health outcomes.

In contrast to previous prognostic models that existed at the time, the PSI was developed as a two step rule. In Step 1, we identified very low risk patients using demographic data that came from -- as well as clinical data that came from the history and physical examination alone.

In Step 2, we took the remaining patients who were not identified in Step 1, and we assessed risk using the Step 1 variables plus a limited set of laboratory and radiographic data that are routinely available at the time of presentation.

So this slide shows the 11 variables that were independently associated with mortality in Step 1 of our rule. They included a single demographic factor, age

greater than 50, five comorbid illnesses, including neoplastic disease, heart failure, chronic renal and liver disease, and cerebral vascular disease, four vital sign abnormalities, including tachycardia, systolic hypotension, tachypnea, defined as a respiratory rate of greater than 30, either hypo or hyperthermia, and the existence of altered mental status.

So these 11 variables can be distilled into a simple three-question algorithm. Is the patient over 50 years of age? If no, did they have any of the relevant five comorbid illnesses? If no, did they have any of the relevant history and physical examination abnormalities, i.e. vital sign abnormalities or altered mental status?

If the answer was no to all three questions, they would automatically be assigned to Risk Class I of the PSI. If the answer was yes to any of these three questions, proceed to Step 2, which is to

quantify risk using the additional laboratory and radiographic factors.

This slide shows the seven radiographic and laboratory factors that again were independently associated with mortality in Step 2. They included acidemia, defined as a pH of less than 7.35, elevated BUN, hyponatremia, hyperglycemia, anemia, hypoxemia, defined as a PO2 of less than 60 millimeters of mercury or an O2 saturation of less than 90 percent based on pulse oximetry, and the existence of a plural effusion on the baseline radiograph.

We had a logistic regression model, and we used the beta weights in the logistic regression model quantify the association between each of these factors and the likelihood of death at 30 days, and what this slide shows is that the points that were assigned to each of these independent factors associated with mortality.

So to quantify a risk score for a

given patient, you take the age in years for males or the age in year minus ten for females, because males had a slightly worse prognosis, and then you serially add points for each of the factors that are present at the time of presentation.

This slide shows the relationship between risk class and the proportion of patients who were dying in the derivation cohort and the two validation cohorts, and I'd like to just walk you through this slide. Is the pointer available? Thank you.

Risk Class I are the patients who are defined by the three-question algorithm alone. Patients in Risk Class II had less than 70 points, and Risk Class III, 71 through 90 points, and patients in general in Risk Classes I through III are considered low risk.

What you can see is when you look at each of these risk classes that there was no significant difference in mortality across risk classes, suggesting that the mortality

that we derived in the initial MedisGroups was well validated in two independent populations. One point I'd like to make is that the cumulative mortality in patients in Risk Classes I, II, and III in the PORT cohort was less than one percent.

One of the things that we were able to do in the PORT cohort is to simulate the effectiveness of using this rule to guide the initial hospitalization decision. We asked, "What if all non-hypoxemic patients in Risk Classes I and II were treated as outpatients, and those in Risk Classes III were treated with only brief inpatient observation?"

In this simulation, this strategy would have resulted in a 26 percent reduction in inpatient care, and an additional 13 percent of inpatients would have had a brief rather than traditional inpatient stay.

There have been a number of standards that have been developed to assess prediction models, and I won't go through each

of these, but I'd like to summarize by saying that the PSI met virtually all standards for prediction rules of prognosis with the exception of really not assessing the impact on patient care.

Since the publication of the PSI, there have been five studies that have assessed the impact of the PSI on guiding the initial site of treatment for patients with CAP. Two were cluster randomized effectiveness trials. There was one randomized efficacy trial, one pre-, post-quasi-experimental trial, and one observational-controlled trials.

These studies enrolled close to 4,000 low-risk patients at 60 sites in four countries. Four studies concluded that use of the PSI increased the proportion of low-risk patients treated in the outpatient setting without compromising patient safety. The fifth trial that was done -- was the randomized efficacy study that was really

designed as a no-difference trial.

For interest in time, I'm going to skip over the next two slides. So what about the methodologic rigor of the PSI as a decision aid? Brandon Reilly published an article in the *Annals of Internal Medicine* that went over four levels of evidence to judge how well a decision aid performs, ranging from derivation to broad validation to narrow impact analysis to a broad impact analysis.

So what's been done for the PSI? In terms of derivation, we've identified 20 independent predictors in 14,000 patients at 73 sites. It was broadly validated in the initial publication in over 40,000 patients from 180 sites.

There was a narrow impact analysis done in our simulation using the PORT data, and since that time, in terms of a broad impact analysis, it's had its effectiveness demonstrated in nearly 4,000 patients at 60

sites in over four trials.

So what are some of the caveats and limitations of using the PSI to guide a site of treatment? These are important to recognize. First, it includes a large number of predictor variables that complicates its use, and the dichotomous nature of the predictors may oversimplify decision-making.

The second, it does not include rare medical complications or conditions that are associated with prognosis and does not consider frailty or psychosocial factors that clearly have an important role in making decisions like which patients should be treated in the outpatient setting versus in the hospital.

It applies only to non-immunocompromised adults, excluding children, pregnant women, immunocompromised including these who are HIV positive, and patients with hospital-acquired pneumonia.

And, finally, it's important --

when using this as a decision aid to help guide the initial site of treatment decision, it's very important to recognize that it's intended to supplement, not to override physician judgment.

So what are some other applications of the PSI? In addition to being used as a decision aid to guide the initial site of treatment decision, it can help physicians quantify prognosis for communication to patients and their families.

It can be used to adjust severity of illness in comparative effectiveness studies and in therapeutic drug trials such as the ones we're discussing at this conference, and it can be used to calculate observed versus expected mortality at the medical provider and hospital levels for quality improvement and quality assurance programs.

So what are some of the summary points? Over the past decade, the PSI has evolved from a prediction rule for prognosis,

specifically a prediction rule for mortality, to a decision aid for the initial site of treatment for patients with CAP. The PSI meets all methodologic standards for such instruments.

Implementation of the PSI in the emergency department safely increases the proportion of patients treated in the outpatient setting. Due to its methodologic rigor, accuracy as a prediction rule, and effectiveness as a decision aid, the PSI has become the reference standard for risk stratification for CAP.

There are citations for those of you who are interested that are attached to these slides, so thank you very much for your time.

DR. GILBERT: So, I think what we'll do, Mike, is take questions for both you and Tom after Dr. Fleming's presentation, if that's all right.

DR. FINE: Perfect. Thank you.

DR. COX: Now I'd like to introduce Tom Fleming, one of our co-chairs and also Professor of Biostatistics at the University of Washington, and Tom will be talking about what criteria should be addressed to do a credible non-inferiority trial and why is this clinically important. It's a very important topic to our discussions today, and thank you, Tom, for joining us and speaking on this topic.

DR. FLEMING: Great. Thank you, Ed. Well, as Ed has pointed out, we've already had a lot of introductory comments about non-inferiority. There's going to be a lot more discussion about it.

What I'd like to try to do is just take a little bit of time here to provide some additional insights into the complexity, into the criteria that need to be addressed when thinking about how to do a valid non-inferiority analysis.

So we have many standard

interventions that exist in many settings that already provide benefit, and yet there is interest in new therapies, experimental therapies, and in some cases because we think they may provide other properties beyond just enhancing efficacy.

So invasive aspergillosis voriconazole may well provide a better side effect profile than amphotericin B. In CAP, a new quinolone could be -- could have a better convenience of administration profile compared to penicillin, and a mother-to-child transmission of HIV, while intensive expensive interventions can reduce mother-to-child transmission, they're impractical or unaffordable where we need them most, in developing countries.

Can we use therapies that are much more cost-effective to be able to achieve broader benefit? Well, if these experimental interventions, in fact, are non-inferior in efficacy to the standards, then these profiles

make them a very attractive alternative option for patients.

So let's take the penicillin new quinolone example. A non-inferiority trial has the obvious appeal that allows us to see head-to-head how this new quinolone is going to compare to penicillin, but we also want to know from this study is the new quinolone effective.

Well, we don't have a placebo in the non-inferiority trial, yet we can get this insight indirectly by comparing the new quinolone to penicillin, as long as we have quality data on how the standard compares to placebo, and as a result, herein lies the challenge and complexity with non-inferiority is having that evidence about the standard.

And so Bob O'Neill, Bob Temple, and others in developing the ICH guidelines have noted that, for penicillin or any active comparator to have appropriate efficacy, it needs to be efficacy that's clearly

established and quantified, and where the efficacy of that active comparator from those historical studies is relevant to the setting of the non-inferiority trial, or in my simplified terms, if we're going to use penicillin as the active comparator, its benefit needs to be substantial, precisely estimated -- ideally from previous randomized trials -- where those estimates are relevant to the setting of the non-inferiority trial against the new quinolone.

So why is this issue -- this is called the constancy assumption, and this is the downfall of so many non-inferiority trials. Why is this so critical?

Well, suppose you were comparing a new experimental intervention against vancomycin, and suppose the result looks similar. The issue is, is your new therapy similar effective to vancomycin or similar ineffective.

You say, "Well, Tom, it's similar

effective." How do you know that? Well, because we have historical data that shows vancomycin was effective in pristine patients, but now we're doing this comparison, let's say, in VRE patients.

Well, it may well be that vancomycin was effective historically, but the critical issue is, in the non-inferiority trial in VRE patients, if vancomycin has much less or little effect, then being the same as something minimally effective is minimally effective.

Well, why is it that if we have historical trials about penicillin, for example, that show it's effective, why do we have to worry about whether penicillin in the non-inferiority trial might be less effective? What are some of the factors that could influence that?

Well, maybe in this non-inferiority trial we have less responsive patients to penicillin, or maybe in the non-inferiority

trial in modern days we have enhanced levels of care that, in fact, attenuate what the effect of the standard penicillin would be, or possibly in the non-inferiority trial there is lower adherence to the standard penicillin, or the endpoints could be different.

And so the reason this is important is that if the standard therapy truly is much less effective than what it was seen in a historical context, then an ineffective experimental will look dissimilar, will look non-inferior, and yet, in fact, it is ineffective, but you would be falsely concluding it was similar effective if you were assuming constancy assumption held when it doesn't.

So let's give a little more insight about how we proceed in all non-inferiority trials but certainly in non-inferiority trials in community-acquired pneumonia. The essence is, we need to define a margin, and we get a relative risk or relative effect of the

experimental against the standard, and we have to rule out that the experimental can be worse than the standard by more than a margin.

Well, the devil is in the details. How do you get that margin? What's a valid margin? So let's illustrate this again. Let's say in pneumococcal pneumonia, where the standard is penicillin, new quinolone is the experimental, and let's say the endpoint is failure.

Failure here, we could define in many ways, depending on whether it's mild, moderate, severe, but it might be a composite of death, of persistent symptoms or breakthrough infections or worsening of symptoms.

So let's suppose that penicillin has a 20 percent failure rate, and let's suppose the new quinolone has a slightly higher failure rate of 25 percent, and with 150 patients per arm, two standard errors is plus or minus ten percent.

And so basically, if we're plotting here on this graph, what is the failure probability on experimental, the new quinolone, against penicillin? We hope we're over here. We hope we're at minus 15, minus ten, where there is a lower rate on the experimental, yet in this setting we had an estimate of a five percent higher rate with a confidence interval that said it could be up to 15 percent higher.

Well, is that upper limit sufficiently low that we can say that this new quinolone is effective or is adequately effective? What's the margin? We certainly would like to rule out that it's meaningfully worse.

Well, one of the aspects in defining the margin is, where is placebo? Where do we put placebo to know whether or not this confidence interval allows us to conclude we're better than a placebo? So we need evidence on the effect of the standard

therapy.

So suppose we have historical data, historical trials that show that placebo actually has a much higher failure rate, a 50 percent failure rate compared to penicillin's 20. So we're estimating it to be 30 percent higher plus or minus ten percent.

So now we can put where placebo is on this graph. Here is placebo, 30 percent higher in its failure rate compared to standard, but we don't know that for sure. That's a point estimate. We have imprecision in that.

Adjusting for the imprecision in that and adjusting for the uncertainty, the constancy assumption, one traditional approach is to say we're going to put placebo here where the lower limit is, so we're only confident or reasonably confident that placebo has a 20 percent higher failure rate than the standard, than penicillin.

Well, you might say, "Okay, this is

great, Fleming, because we can rule out that the new quinolone is more than 15 percent, so we can establish efficacy." Well, part of the problem is, if you have an effective standard like penicillin, why is it enough to find something new that's better than nothing?

A tradition that's emerged is you want to at least be able to preserve some fraction of what it is that the effective standard provides, and a frequent margin, then, is based on preserving at least half the effect of the standard, and that would then yield a margin of ten percent.

And so essentially what we would need to have is a confidence interval that rules out that we're more than ten percent worse in mortality to be reasonably confident that we're preserving at least half the effect of penicillin.

But there's another consideration that should come into account in this margin, and that is, is a patient comfortable that

your agent could be losing ten percent, and if this were mortality, would you, in fact, achieve or be willing to take a ten percent higher mortality rate for a more convenient administration? Now the temptation to say, "Sure, we want this big margin," but I always say, "Turn it around. Turn it around."

Suppose you had a 30 percent mortality, and you could reduce that mortality to 20 percent. Would you be off to the FDA filing to get an approval for superiority because clinically you've reduced the death rate by one-third? You bet you would, and so if it's clinically meaningful to provide a ten percent improvement in the failure rate, why is it clinically acceptable to allow more than or up to a ten percent loss?

So bottom line is clinical relevance also enters into the definition of the margin. So the reduction in efficacy that we're allowing needs to take into consideration these other issues, and these

need to be pretty powerful to allow for a meaningful reduction in efficacy.

In essence, this margin needs to be sufficiently conservative that you can reasonably conclude that this agent is preserving at least half the effect of the active comparator that, in effect, is effective, and that we're not allowing for a clinically meaningful worsening of outcome.

John Powers already raised the question what is the conclusion if we do this non-inferiority trial, and it's positive -- you rule out the margin -- what's the conclusion? That the new quinolone is at least as good as penicillin, that it's not worse than penicillin? These are what you see. This is what is commonly stated as the conclusion.

Well, let's suppose you actually did a bigger non-inferiority trial, and in that non-inferiority trial, once again, the quinolone is five percent worse than

penicillin, but now with its bigger size, it's plus or minus three percent for two standard errors.

So what does this graph now show? This graph now shows that the new quinolone still is five percent worse by estimate, but we know it's not more than eight percent worse. Therefore, we have established non-inferiority, but oh, by the way, it is two percent worse, so it is inferior.

Okay, this analysis establishes that the new quinolone is non-inferior to placebo while proving it's inferior. I'm perfectly comfortable with that. It's true.

How can you be comfortable with that paradox? You're comfortable with it because non-inferiority doesn't establish that the new quinolone is not worse than. It establishes that it's not unacceptably worse than.

You can be worse without being unacceptably worse, but it means this margin

does need to rule out all differences that would be unacceptable, so if you want to take a big margin to do a small trial, how in the world can you argue that anything less than 20 percent or 15 percent is not clinically meaningful?

So the bottom line is no, these aren't the conclusions of non-inferiority. Marketing people won't want to hear this, but if you establish non-inferiority, what you can legitimately say is the new quinolone is not meaningfully worse than penicillin. That is what you have established with non-inferiority.

Quality also matters, quality of the trial conduct, and once again, if we go to the ICH guidelines, any trial needs to have high quality. A non-inferiority trial has an even higher bar. Why is that? Noise in a trial, missing data, non-adherence, poor quality conduct leads to lesser detection of differences.

In a superiority trial, well, you can say, "If I showed an effect a fortiori, I would have shown even more of an effect if I had a clean trial," but in a non-inferiority trial, if there really is inferiority, you're going to miss it because of poor quality, and so, as ICH says, it's critical in a non-inferiority trial to have high levels of adherence, high levels of retention, et cetera.

We've already talked about this issue of capturing all of the outcomes. As ICH says, it's especially important to minimize loss-to-follow-up and missing data, as Bob Temple has already pointed out, but if you have an absence of the targeted microbial, you're not going to benefit those patients, and so there is the temptation to pull those out, but if only half of your patients are found to have the targeted microbe, and you leave those half out, what you have to recognize in benefit-to-risk is benefit-to-

risk isn't fully defined by benefit-to-risk in the targeted half, because you also have all the other half that carried that risk.

I think it's very poorly advised to do a trial where substantial fractions of people will be identified later to not have the targeted microbial, but these kinds of missingness also are frequent and highly problematic. Leaving patients out of the analysis because they had adverse events, they didn't take the therapy, they didn't perceive benefit destroys the integrity of randomization.

Now, all right, so Fleming, you say you need a rigorous margin in order to be able to assess efficacy, but isn't it true that if you use a rigorous margin, you're going to have to have an obnoxiously large sample size? Fact or myth?

So let's continue to look at this situation where the standard, let's say, has a 20 percent failure rate. What I'm plotting

along this axis is the experimental minus the standard failure rate.

If, in fact, you allowed a margin that allowed up to a ten percent increase in this failure rate, where you only had a 20 percent failure rate in the control, you're having to argue that a relative 50 percent increase in the failure rate is okay before it matters to patients.

A 20 percent margin basically says a 25 percent, 50 percent, and 75 percent relative increase in failure rate is okay. You just can't have a doubling in the failure rate. So does it, in fact, meet clinical common sense that a margin that high could be defensible?

Well, so let's look at why a rigorous margin doesn't necessarily require a huge sample size. So again what I'm plotting along this axis here is the failure rate on experimental versus standard, so we'd love to be out here at minus 12, where the

experimental is reducing the failure rate from 20 percent to eight percent. It's 12 percent better.

If you're doing a superiority trial, then essentially, you need to have a point estimate sufficiently negative and a confidence interval ruling out equality to establish superiority. If, in fact, your experimental is 12 percent better than the standard in the failure rate, then with 340 patients you have 90 percent probability of a positive result, 90 percent power to rule out equality.

Well, that's great, no controversies, but lots of antimicrobials aren't that good, and if we held the bar to being that good, we may have a hard time finding new interventions.

So the idea is let's look at non-inferiority, and let's be lenient here. Let's let a non-inferiority margin be 15 percent, and essentially then a positive trial would be

a point estimate, a confidence interval, that would rule out your 15 percent worse. Well, if the experimental truly is the same as standard, then with 300 patients you have 90 percent power to rule out your 15 percent worse.

The problem here is even if you're ten percent worse, you've got a substantial probability of ruling out your 15 percent worse, so with such a lenient approach, you have a substantial probability of approving agents that are a lot -- clinically, a lot worse.

All right, so we fix that by using a ten percent margin. Now, if you truly are the same, experimental and standard are truly the same, now with 672 patients you have 90 percent power to rule out the margin. You're ten percent worse, but you say, "Aha, Fleming, I told you. See, you used the rigorous margin. Now you have to have a really big sample size. Now it's 672."

The problem with this logic is, why is it that if we don't think we're a whole lot better so we can prove superiority, that the only other thing we could be is the same? Isn't there such a thing as being a little better?

What if you're three percent better? You're not going to be enough better to show superiority in a practical trial, but if you were three percent better, couldn't you rule out a rigorous margin with a reasonable sample size? And you're right. You're absolutely right.

If you're three percent better, if the experimental has just a three percent better failure rate than standard, then with a more attractive sample size you do have 90 percent power to rule out a ten percent margin. You even have 70 percent power if you're the same.

So rigorous margins don't make it difficult to establish benefit of agents that

are just even a little bit better. Yes, it makes it a little more difficult to establish benefit of agents that are the same, but there is not as much downside to public health by having a little bit harder time to establish benefit of agents that are just the same.

Now, question. What are these green asterisks? What are the green asterisks? The green asterisks are the least favorable estimates that would allow you to get a positive conclusion.

So you have to be estimated to be seven percent better to show superiority. You could be -- you have to be at worse two percent worse to get non-inferiority here.

The problem with this lenient margin, you could be estimated to be six percent worse. That's a 30 percent relative increase in failure rate, would be enough for non-inferiority, when you're using such a lenient margin. Using those lenient margins that give you the ability, even when your

estimate says you're a fair amount worse to get an approval, leads to bio-creep.

What's bio-creep? Well, you don't just do non-inferiority once. If you now do non-inferiority, then non-inferiority and non-inferiority and then non-inferiority, how long before I have no clue what I have? Okay, that's not hypothetical.

When I was serving on the Anti-Infective Drugs Advisory Committee in the 2002 February meeting, I brought up this example that was presented to the Antiviral Committee in 2001 that was looking at voriconazole for empiric antifungal therapy of febrile neutropenic patients, and we had a series of non-inferiority trials.

First it was amphotericin B, then ambisome to amphotericin in non-inferiority then vorciconazole to ambisome in non-inferiority. The endpoint was this composite failure endpoint. Death, breakthrough fungal infections, persistent fever. Any of those,

you were a failure.

Okay. Well, first of all, these were uncontrolled, small trials. This did, at least, have a point estimate that was similar, but the voriconazole estimate was about six-and-a-half, seven percent worse, and by the way, the failure rate in ambisome here was different from here. It used a different endpoint.

Well, this actually showed that you had a significantly higher failure rate on voriconazole, but the upper limit was 12 percent, so if we used a 15 percent margin, we're okay; right? Well, the Antiviral Advisory Committee said, "No. No, stop."

I have no clue from this data what voriconazole really is doing, but if they had approved it, what would your fourth generation antifungal compare to? I'd use voriconazole for my active comparator, and let's use the 15 percent margin again. How long before a placebo isn't highly likely to succeed in that

scenario?

So many of us would say in the anti-infective setting when we do non-inferiority trial on non-inferiority trial in acute otitis media and acute bacterial sinusitis and acute exacerbation of chronic bronchitis, what do we know? What do we really know about efficacy of those interventions?

So some summary comments. A successful non-inferiority trial does not lead to the conclusion that you are as effective as the standard. It simply allows you to say you're not unacceptably worse. Okay, therefore it becomes critical to define what is the smallest difference that is unacceptably worse.

Therefore, margins should not be based solely on a statistical calculation. Margins should not be based on, "Well, let's see. If we have an 80 percent cure rate, we want a 300-person trial. What's the margin we

could rule out with 90 percent power? Twenty percent. Aha, that's my margin."

No, that's not how we define a margin. First of all, the margin should be smaller than differences in efficacy that patients and care givers would consider clinically relevant, but furthermore, the margin isn't just an issue that is based on clinical judgment.

You need scientific data to also establish that you truly are ruling out placebo and even more so that you're preserving half the effect of an effective active comparator. Bio-creep can be avoided with rigorous margins, and rigorous margins don't lead to huge sample sizes for interventions that would be at least moderately better.

Really quickly, non-inferiority trials on surrogate endpoints I often call my worst nightmare. Non-inferiority trials are already bad enough, but you're trying to come

up -- let's say it's a clinical endpoint of mortality.

At least we have a shot at defining what's the loss in mortality that we care about. Tell me how much loss in an antimicrobial effect we have to have to translate to that much loss in mortality: an incredibly uncertain, complicated issue.

Non-inferiority trials share many of the dangers of historically controlled trials, and as a result, they should be avoided if at all possible. It is an extremely unfavorable way to try to establish efficacy.

Any way that you could do superiority would be far superior in terms of understanding the effect, and, in fact, a recent article in *Lancet* goes one step further. This article that just came out in *Lancet* said non-inferiority is unethical.

The argument that the authors are giving here is, if you're comparing an

effective standard, and you do a randomization where you have a half a chance to get that, and you have a half a chance to get another agent that, in fact, we hope is as good as, but could be clinically meaningfully worse, why is it to your advantage to be randomized to an agent that you hope is as good as, but could be clinically meaningfully worse?

It does point out that if you're going to do that, that other agent sure better have some other really tangible good things that will motivate the patient to be randomized to something that isn't more efficacious and could be meaningfully worse in its efficacy.

You need, though, to have substantial magnitude for your active comparator, precisely estimated ideally in randomized trials, where those estimates are relevant to the setting of the non-inferiority trial, a condition that is frequently not present, if not highly likely, not present.

So the bottom line is, again going back to the wisdom of the ICH guidelines, the determination of a margin in non-inferiority trial is based on both statistical reasoning and clinical judgment and should reflect the uncertainties in the evidence on which the choice is made and should be suitably conservative.

So if you don't have an agent as an active comparator that's highly effective, precisely known in randomized trials, where those estimates are relevant to the setting of the non-inferiority trial, you're not going to be able to come up with a non-trivial margin, and if you don't come up with a non-trivial margin, then practical common sense would say if you want to know whether or not you are providing benefit, we need superiority trials, not just placebo trials. That's one example, but add-on trials or other superiority trials, if we want a reliable estimate of benefit to risk.

Dave, back to you.

DR. COX: Thanks, Tom. And questions for Tom?

DR. GILBERT: Tom and Mike both, Dr. Fine.

DR. COX: Michael Fine, too, and Bob.

DR. GILBERT: Michael should be here, too.

DR. TEMPLE: If you know the effect of an intervention is quite large, and you choose which -- so that your non-inferiority margin to show any effect would be pretty big, but you choose on clinical grounds a much smaller margin, then I think the need to be very precise about estimating the real effect of the active control is diminished.

So, for example, in some form of pneumonia -- I've got another question to follow that -- if you know it's somewhere between 30 percent and 60 percent effective, you don't really have to know which it is if

your margin is ten. So that seems worth remembering.

In at least some antibiotic settings, you know, urinary tract infections, things like that, my impression, not knowing anything about it, is that the non-inferior -- the clinically derived margin is much smaller than the actual effect of the drug, so that seems worth remembering.

My second question goes to Dr. Fine, and that is, some of the problems that Tom described are based on not knowing who the population is now compared to who the population used to be.

So what I'm wondering is if, whether you could use that scale to assure that at least -- it's ugly to say this -- at least some people in your trial will die, and then that would be confirmed, because they died despite good treatment, if I understand your findings.

Maybe a criterion for CAP is a

population sick enough so that even with optimal treatment, a reasonable number of -- sorry, it's ugly to say it -- some of them died, because that's the way it is. Maybe that provides assurance that you have a population that resembles the population in which you determined these drugs were effective.

So that -- do you understand my question?

DR. FLEMING: Yes.

DR. TEMPLE: I mean, if you use one through three, how are you going to know how to compare it with the past? You just won't, but if you have a nine or ten or 12 or 15 percent mortality, that does suggest it's a little like the populations where you gained the impression that antibiotics work, and it might be reassuring, so that almost sets an ugly standard for us to insist that the population studied have some mortality.

DR. FINE: So I'm not exactly sure I

totally understand your question, but I think that one of the things that the PSI does very well is accurately risk-stratify. So you can say with a fair amount of confidence that if you're focusing on Risk Classes I and II or even I, II, and III that choosing mortality as an outcome measure would probably be a very difficult outcome measure to have for a trial, because to show, you know, differences on an expected mortality of one percent, from one percent to two percent, are going to take thousands and thousands of patients.

So I think that you can certainly use it to define the population and then to use that population to decide what is the appropriate outcome measure to be looking at for antibiotic trials.

DR. TEMPLE: I had something slightly different. One is to maybe choose people by that standard on who to put into the trial, but you then would be testing your assumption by insisting that, nasty as this

sounds, the outcome include a mortality rate. I mean, if nobody in the trials died, then you didn't put the population into the trials where you got your impression that antibiotics work.

DR. FLEMING: So --

DR. TEMPLE: You would insist --

DR. FLEMING: Right.

DR. TEMPLE: -- that the population is --

DR. FLEMING: So I think your question is clear, and I think the issue in answering your question is, distinguish between a prognostic factor, a covariate that's a predictor versus an effect modifier.

What Michael's work, impressive work, I'd say, has done is it's shown in a validated way that we have the ability to identify predictors. We have the ability to assess prognosis, and that's relevant, as you say, because you want to get a population at sufficient risk that you're going to see

events that you're trying to assess.

But what these analyses don't do is they don't tell us what the effect modifiers are, so you're going to be able to use the PORT score to define which patients are at higher risk for mortality, but that's a different question than telling me in which populations is penicillin going to have a big effect on mortality, and it's not the case that, let's say hypothetically, if males have a higher death rate than females, it doesn't follow that treatment effect in males is higher than treatment effect in females. That's the issue I really want to know if I'm going to change this margin.

You're correct, Bob. If I change the population of patients that are in my non-inferiority trial than in my historical trials of penicillin, it might be that those historical trials of penicillin were full of patients that were really sensitive to and benefitted by penicillin, and if I now go to a

new population where this is an effect modifier, i.e. a population now where patients are less benefitted by penicillin, then I have to worry that this margin is even too conservative, because standard placebo actually may lie right here, and that was the example I gave with VRE.

Vancomycin might have a really big effect on the cure rate, but in VRE patients it may have a trivial effect on the cure rate. The PORT score is going to tell us what are the prognostic factors for cure rate. It's not going to tell us what are the effect modifiers for penicillin's effect on cure rate.

DR. TEMPLE: That's true, but if in the past studies from which you derived your estimate of the effect of penicillin, there was a mortality reduction from 50 percent to 30 percent --

DR. FLEMING: Yes.

DR. TEMPLE: What you know is that

despite effective therapy, there was still a mortality in the population that's relevant on which you based it.

DR. FLEMING: Yes.

DR. TEMPLE: That to me means there should still be a mortality in the new trials. It shouldn't go to zero, because that raises the question of whether they were at risk at all.

DR. FLEMING: Yes, of course, there are many different things that can change, and how many of those that we know and we can put into our model I always say is the tip of the iceberg. The bottom line is --

DR. TEMPLE: These are recent studies, though. Under current treatment paradigms, there still is a mortality side.

DR. FLEMING: And the more recent and more relevant these studies are to the context of the non-inferiority trial, the more confident I'm going to be about where placebo lies.

DR. TEMPLE: But that's the trouble. The placebo-controlled trials of pneumonia are old. We know that.

DR. FLEMING: That's the trouble, yes.

DR. TEMPLE: But what I'm still asking is, if you pick a population which on treatment still has a mortality, aren't you at least moderately reassured that the past data are relevant to the population you put in your trial? That's what I'm asking.

DR. FLEMING: I would argue more, Bob, I'm moderately reassured to the extent that I can argue that the patient characteristics are similar, the supportive care is similar, the adherence to the active comparator is similar, and the definition of the endpoint is similar, and if all those are similar, then I'm going to have a greater confidence that the estimate of the active comparator effect from the historical trials will apply, and to be truthful in this

setting, it's a real stretch to argue that many of those factors are true.

DR. COX: Thanks. Bob?

DR. O'NEILL: Yes, I have a couple of questions related to -- probably -- the implications of these presentations on study design. Tom makes the point or several points that, first of all, the non-inferiority design relative to other choices, superiority in particular, in many ways is a second-class citizen.

You should not do that design if you have another choice, because there are too many risks in doing it and coming up with a conclusion that you think you have an effective product when you really don't.

And there is a real distinction here, even in the literature, between use of this design when you already have two known effective agents, and all you want to know is whether these two agents are reasonably like each other.

The conclusions from that design are different from the conclusions of what we're talking about here, which is, one of these is ineffective, and if you make the wrong decision, you've really made a wrong decision. You have essentially allowed an ineffective agent to be on the market, so that's why this design, just from its basic principles, is, you've really got to get everything right at the design stage.

The point I'm going towards here is, to my way of thinking, one of the major problems with this design in the anti-infective area is the inability to identify people who are going to benefit at the entrance criteria level.

So you get a mixture of people who are either -- have to be thrown out after the fact because they don't have the bug, and that's not a good idea. That's bad practice. If you don't have to do that, it would be a good idea not to do it.

So we've heard some ways of being able to "enrich the population" by screening early and not being in that game, and we've heard that from the Swiss work that's going on, and we've also heard it from your -- essentially -- prediction tool.

So I guess my question would be how do you see both of these, both the Swiss presentation and what's been going on there in terms of the interesting thing is you get the same outcome if you treat for five days or four days. We heard a comment from the floor saying there is very little data on these studies being done in, I guess, hospitalized or ICU patients.

Your categorization essentially has a lot to do with whether you put somebody in the hospital or whether you don't put them in the hospital, whether you treat them outpatient or inpatient.

This has a lot to do with what the future designs are in terms of maybe

stratifying on that, maybe taking some of the Swiss thinking into play, and actually designing superiority trials which show a difference. Tie them into the John Powers idea of a dose, two different doses.

So I think there is opportunities here to, at the very least, change the enrichment strategies for all these trials so you don't get in the game of diluting your signal, because you've got a mixture of populations, some of whom don't have bacteria but have a virus, and secondly, some of whom are resistant, and some of whom aren't, but you don't know, and is there a way --

And all of those things, mixture populations who have differential response, are really bad for non-inferiority trials, because it's a source of noise that drives you to the null, and that's where the problem is, and the only solution is to upsize the trials, and nobody wants to do that, and that's not even satisfactory.

So I guess my question is, to both of you, how do you see, with your presentation and with the -- what Dr. Niederman's presentation about what's going on in the Swiss clinical trial thinking, to actually dramatically change the entrance criteria to trials as to what the current practice is, because I think that alone is a huge benefit, that alone, from current strategies.

DR. FINE: Let me take a crack at this first. I think that I can probably comment less on the calcitonin literature, and I'll let Dr. Niederman make comments on that.

With regard to using the Pneumonia Severity Index or some objective measure of pneumonia severity in defining clinical trials, I think that it can be used to define homogeneous subsets of patients that share similar risk for a given outcome, and the outcome of interest in the PSI is mortality, so inasmuch as those individuals who are interested in developing clinical trials for

antibiotics want to define severity based on risk of mortality, I think it could do a very good job in achieving that goal.

Although tempting -- it's tempting to use site of treatment, I think that there are some inherent flaws in using site of treatment as a proxy for severity of illness, because there is so much variability in physician decision-making with regard to site of treatment, not only for the home-versus-hospital decision, but Dr. Wunderink brought up ICU patients.

There is also a fair amount of variability from hospital to hospital and provider to provider and medical system to medical system of who gets into the ICU and who doesn't get into the ICU, or is it -- are they patients who have frank respiratory failure and sepsis and hypotension, or are they placed there because the hospital has a shortage of nurses, and they can't get adequate observation on the floor?

So it's a little bit of a long-winded answer, but I think that the PSI has the greatest potential to make a contribution if those interested in defining criteria are interested in defining severity of illness based on objective stratification of risk of mortality.

DR. O'NEILL: The reason for my question is that it only -- if this was a cardiovascular trial, and the idea would be you enter higher risk patients, because the probability event is higher.

So that's really the strategy that you're talking about, but that doesn't solve the other antibiotic, anti-infective problem which is entering people who don't have the infection, so you've got the other piece of it that you've got a mixture population.

DR. FLEMING: So let me try to respond again, because you're right, and this is a distinction that's often missed. When we define the enrichment population, one approach

that we take, because this is where we have the data, is who are the patients at higher risk for the endpoint of interest, because any data that exists in this setting is informative about that question.

What is far less informative, though, or what's far more difficult to determine isn't who are the patients that are more likely to have an event. Who are the patients most likely to benefit from the therapy?

Bob Temple gave an example in advanced breast cancer of Herceptin. Basically, understanding the mechanism of Herceptin stated that you want to use people that had high levels of HR2 over expression.

Therefore, we enhance the sensitivity, not because people with high levels of HR2 over expression were more likely to die, but because those are the people likely to benefit from the mechanism of action of the intervention. So we need much more

than just who is the prognostically high risk category. We need the insight about mechanism of action.

There is a whole lot of discussion about that out there nowadays looking at targeted therapies. It's really the main message in oncology. It's a great idea.

It's incredibly difficult to understand all of the intended and unintended mechanisms of intervention to be able to really reliably understand, in advance, who are those people most likely to benefit, but if you can, that's how you get an enriched study, but then your label is correspondingly restricted.

So with BiDil in heart failure, when they did a registrational study in blacks alone, because that was their enriched population, their label is blacks alone, so if you do think you have an ability to enrich, and you do a very targeted population, you're only assessing benefit to risk, then, in that

population.

DR. COX: Dr. Niederman and then Dr. Echols.

DR. NIEDERMAN: Let me just amplify this last point, because I think it is really important to understand. You're trying to enrich patients who are going to benefit from an antibiotic trial, and I think that that's been stated so many times as a misunderstanding here, that PSI --

I think Michael said it well. It's a predictor of mortality. It is by no means a measure of the ability to respond to antibiotics, and it's by no means necessarily a predictor of the severity of the pneumonia itself, and the way you've got to understand this is, if you're a 75-year-old male, and you've got a history of prostate cancer, tomorrow when you wake up feeling great, you're PSI IV, and it's got nothing to do with whether you're going to respond to an antibiotic in a clinical trial.

It just means that if you get a pneumonia, you might die from it, and that's a very different statement. So I think you have to understand the distinction.

I'm not trying to push or not push the issue of procalcitonin, but at least when those trials were designed, they were looking at a relevant endpoint. They were looking at the benefit of antibiotic therapy and the withholding of antibiotic therapy, and that's why, conceptually, that can enrich the population in a very different way than you could enrich the population by using PSI scoring.

DR. FINE: I guess, just as a follow-up comment, the PSI score is really predicated on having the diagnosis of pneumonia already established, so it's really not going to be helpful in terms of saying who has the disease entity of interest versus who doesn't, who would be even eligible for an antibiotic trial.

DR. COX: Thank you. Dr. Echols?

DR. ECHOLS: Yes, thank you. Actually, my thoughts were a little bit along the way Michael is going, but first, you know, my concern is this sort of emphasis on mortality as an endpoint, and just to go back to where we were at eight o'clock this morning, describing the discussion and mild to moderate CAP, which constitutes 80 percent of treatment courses for CAP and diagnoses for CAP, mortality as an endpoint is really not feasible, and we can't take a drug that's necessarily approved for severe infection and then translate it as we discussed earlier, that it'll also work for mild to moderate.

So the other point is that the PORT scores, looking at mortality, and I -- this is more of a question than anything else, but was there effort made to determine whether that was attributable versus associated mortality? Because, as I've conducted clinical trials and had many deaths, and I have to write up a

detailed precis and review of every single death in every trial, and what I've been impressed by is, particularly with patients that die in oral therapy drugs, that the mortality has nothing to do with the infection being treated, that the mortality is related to underlying cardiovascular disease or thromboembolic disease or CNS disease, and to even consider mortality, even in severe pneumonia as directly related somehow to the treatment and the drug-bug interaction and the treatment of an infection, I don't see the data that really supports that mortality would be attributable to the infection.

DR. FINE: So that gets at the issue of all-cause mortality versus disease-specific mortality, and one of the things that we did as part of the original pneumonia PORT cohort study is we took a look at each and every one of the causes of death, went into the chart, traced out of all the circumstances that intervened between the time of patient

presentation and death to try to assign whether the death was pneumonia-related or whether the death was not pneumonia-related, and I can't recall the exact numbers, but only about half of the deaths were specifically pneumonia-related, so you're right. About at least half of the deaths are related to underlying comorbidity that had nothing to do with the actual infection based on a strict cohort review.

DR. ECHOLS: And that was the cohort that just had hospitalized patients, or did that include --

DR. FINE: No, that was inpatients and outpatients.

DR. ECHOLS: That was the -- okay.

DR. FINE: But there were only seven patients in the 944 outpatients that died, so most of those patients ended up being treated in the hospital.

So there are good data looking at what the causes of death are. As I recall,

there was no good correlation between severity of illness and whether they died from their pneumonia versus a comorbidity.

DR. ECHOLS: Okay. Thanks.

DR. FLEMING: And just a 30-second follow-up comment. Maybe this is obvious, hopefully. Whether you're talking about severe CAP, where mortality might be an endpoint, or whether you're talking about mild or moderate CAP, where resolution of symptoms would be an endpoint, these principles that we talked about are the same in those two settings.

Now, whether they're satisfied could differ. We may be able to justify the principles as being valid for mortality and severe but not for resolution in mild to moderate, but the point is the principles are the same, independent of the disease setting or the endpoint used.

DR. COX: Thank you very much for the comments and questions. Now I'd like to

move on and invite Dr. David Gilbert to present. Dr. Gilbert is Professor of Medicine and Chief of Infectious Diseases at Providence Portland Medical Center, and he'll be talking to us today about clinical endpoints of therapy to include patient-recorded observations. Dr. Gilbert.

DR. GILBERT: So the original program listed Jack Edwards as giving this presentation on clinical endpoints for mild to moderate CAP. Dr. Edwards sends his apologies. He's got a health problem, and so at the last minute, relatively the last minute, he asked me to substitute for him. I'll try to do the presentation justice.

These are my bad habits, if you will. The plan is to briefly present a historical perspective. As you've heard from others, we have to base a lot of our judgments on the work of our pioneers in the early days of infectious disease, and then discuss current suggested approaches to quantifying

endpoints or outcomes. I'll go back and forth on that terminology, and I've had help from a lot of colleagues in the presentation.

So from the historical perspective, I was attracted by this paper by Petersdorf, Cluff, Hoeprich, and others in the Bulletin of the Johns Hopkins Hospital in 1957. Drs. Petersdorf and Cluff and Hoeprich were sort of the pillars of the foundation of infectious disease as a specialty, and they conducted a prospective randomized double-blind trial of what we would consider low-dose penicillin today and intramuscular every 12 hours for seven days or until afebrile for 48 hours, and then you either got aspirin or a placebo tablet. Didn't say if it looked exactly like aspirin, but a placebo tablet.

The interesting thing is that patient symptoms were evaluated independently by two MDs who were blinded to the therapy, and John Powers showed some of the difficulty in reproducibility of various MDs eliciting

symptoms, and those of you that knew Dr. Petersdorf and his personality, I would have loved to have watched him collect this data from a patient.

Dr. Petersdorf was always in a hurry, and I suspect he took about ten seconds to assess symptoms in the area of general symptoms, appetite, cough, and pleuritic pain and then had sort of the beginning of a Likert scale grading the severity of those patient symptoms.

And, of course, this relates to some of our modern concepts of patient-reported observations, and then they reported these results as a percent of symptom, becoming asymptomatic over a five-day period, and obviously there was no difference detectable between the aspirin and the placebo, but the effects became evident fairly quickly.

Now this doesn't apply directly to our mild outpatients. These were inpatients,

and yet they were able -- at least the authors thought they were able to get reasonable data from inpatients who obviously were feeling quite ill.

The temp curve, on the other hand, looks like modern art, so they didn't do too well putting this into a time-to-response parameter, and some of the more modern studies, which I'll get to momentarily, have done just that.

So what's happened in more modern times is this patient-reported observation as an endpoint or an outcome has utilized the techniques of colleagues in sociology and psychology, psychometrics, the branch of psychology that designs, administers, and interprets quantitative tests used for the measurement of psychological variables and then have adapted those techniques to a clinical setting and then another phrase that I wasn't as familiar with, clinometric, which is similar, but it's assessing symptoms,

signs, and laboratory results by scales, indices, and other quantitative instruments.

Of course, if you then go to that literature, you get into a lot of terminology that we clinicians, even academic clinicians, are not terribly familiar with, but obviously terribly relevant.

Reliability. Whatever questions you ask have to be stable over time, reproducible between different observers, have to be valid to the extent to which the endpoint measures what is intended.

Responsiveness, detection of the complications that we want to know if the complications are present, and acceptability to all of the users, and there is a nice review of all of these in much greater detail than I have time to present in this *Lancet Infectious Disease Review* article in 2003.

So we've heard a lot of discussion about endpoints, so I'm going to go now to some of the classical endpoints and then come back to

the patient-reported observations.

As you're heard about eight times now, most patients with mild to moderate community-acquired pneumonia don't die, so that's not very good for an endpoint, and Dr. Fine described those numbers.

The mortality of -- one other paper I found interesting from the *Annals of Emergency Medicine*, the mortality of outpatient CAP very low, as already mentioned, but if those patients subsequently developed a complication and had to be admitted, then there was a tenfold increase in the mortality rate. So, obviously, you'd need a huge sample size so mortality is an insensitive endpoint or outcome measure.

For patients that get into the hospital, not directly applicable to our outpatient patients, the length of stay is often quoted, but, gee, that's influenced by clinician's practice style, the need or pressure on hospital beds and the efficiency

of discharge planning.

There is variations in length of stay between a hospital without validation -- without variations in outcomes, so the length of stay hasn't correlated with variations in outcomes. There is a phrase, time-to-clinical-stability, but that again isn't applicable to outpatient therapy.

I'm very envious of our colleagues in the viral field, because they have to deal with clearer endpoints in that they can measure the viral load and follow the viral load very easily for HIV, Hepatitis B, and Hepatitis C, and we have the CD4 counts as accepted surrogate markers and so forth. Maybe we're edging that direction, as you heard from Dr. Nolte's presentation, but we're clearly not there yet.

In terms of the micro biologic response, we certainly want to try to find out the microbiologic etiology, especially if there is bacteremic patients, endocarditis

patients, and so forth, but those sorts of issues don't seem applicable to outpatients.

We've already heard that the current methods aren't terribly good at detecting etiology, but there is hope on the horizon. So the term microbial eradication or presumptive microbiologic eradication makes no sense, as Dr. Powers brought up.

We need the chest x-ray to enroll the patient, so it's the gold standard for diagnosis. It doesn't seem very useful as an outcome measure. Even though there is sparse data on outpatient community-acquired pneumonia in terms of the x-ray, there is good data on inpatients.

So one study, 288 patients with severe community-acquired pneumonia, by day seven, 25 percent of the patients had an improved chest x-ray, 25 percent, but 56 percent were clinically improved. I know I'm not supposed to say improved. I'm just quoting the authors. By day 28, 53 percent of

the patients had improved chest x-ray, so, far less than 100 percent, but close to 80 percent were clinically cured by day 28.

So it's not a useful or a practical endpoint, which gets us back then to patient-based outcomes. So the idea is to capture the features of outcomes that are of importance to patients.

A lot of these are subjective symptoms that can only be assessed by the patient. We apply these tools of psychometrics, as I described earlier, using numerical scales.

So the questionnaire, the questions that are asked, are not things that you can create on the back of an envelope chatting with colleagues. They have to be documented as reliable, valid, responsive, and actually several other criteria.

So the best example, and it may be to date, and others will correct me. The only example is a questionnaire instrument that was

developed at the School of Public Health in London, a community-acquired pneumonia symptom questionnaire, and it was developed for use as part of an endpoint assessment in a prospective randomized double-blind study that compared oral moxifloxacin to either oral amoxicillin or clarithromycin over a 14-day period.

Sixty-four centers in 13 different countries, 556 outpatients, and the questionnaire, and this I found very impressive, was developed in English but then was translated and successfully utilized in 12 other languages.

So the interviews were conducted by phone or face-to-face and completely standardized. Literally every single word that the interviewer asked was scripted, if you will, at three time points, and this was critiqued by John Powers earlier, at study entry day three to five, during therapy and at the end of therapy, and so in hindsight we

would have liked to have seen more front-ended assessments of the endpoints.

They utilized 18 community-acquired pneumonia-related symptoms, cough, sputum production, dyspnea, chest pain, using a six-point Likert scale, and I'll get to the scale in just a moment.

All of the items were tested for the psychometric criteria. Acceptability, reliability, validity, and responsiveness were the major criteria that were utilized.

And I just scanned in the top half of the scale. It goes down to 18 items, and as it says at the top, "Read to each patient. Read each item to the patients and circle the number that corresponds to how the patient has been bothered by the symptom in the past 24 hours." So Likert scale from zero to five, so five times 18, so 90 would be the maximum point score, if you will, if you circled the five on every single item.

And then this questionnaire was

applied in this study, and these are the results. So 233 patients, 244 patients standard therapy and a very high success rate. Ninety-three percent, I think, was the lowest.

Up to 99 percent of the individuals enrolled in the study completed the interview process, and it's not surprising the scores were nowhere near 90, because these are outpatient mild to moderate pneumonia, and the good news is that the scores lowered during therapy, which most of the -- all of the patients are getting better and that the standard deviations are reasonably small.

So what has been the use of these patient-reported observations in clinical trials? Well, I showed you the one study where it was validated. In the next slide or two I'll talk about a Gati. versus clarithro study that was reported in AAC in 2006.

It was similar, but they used a different questionnaire instrument. It wasn't

the same instrument, and I'll ask -- I should have asked in advance, but I'll ask Ed and colleagues from the Agency if a patient-reported observation endpoint data set has been part of any new drug application to date.

This is the results of that gatifloxacin versus control study using a different instrument and the average symptom score on various days, and they did do a Day Two response and a Day Five response, so as we were discussing earlier, they did measure the early time points using a slightly different questionnaire.

So the FDA has seen the potential value of this, and on February 2 of last year the -- well, two years ago now -- the FDA published a draft guidance for industry patient-reported outcome measures use in medical product development to support labeling claims, and then a symposium was held at the Mayo Clinic. The results of that symposium were just published in *Value in*

Health in 2007 in Supplement 2, which I think just became available, so the symposium was six to 12 months ago, I suppose.

So another -- so patient-reported observations are clearly an evolving, important endpoint. Other endpoints are time-to-event, well known examples, time to normalization of baseline elevations in temperature, time to normalization of the white count and the differential white count.

My little parenthetical comment here is that we've had trouble at our institution and similar problems at other institutions with the automated analyzers that do the CBCs, and most clinicians and, I think, clinical trial investigators are unaware that those instruments are set to detect a band count of between 18 and 20 percent.

So you can have patients that have a significant bandemia, or what most people would consider significant, 15 percent, say, and you would never know it unless you were

doing routine manual differentials.

This is back to the Gati. versus clarithro, just to show the confusion that can occur. This was time-to-event, and all of this (sic) are the subjective symptoms and the only time-to-event objective response that I'm referring to on this whole long list is fever.

This was another trial, moxi versus amox/clav, where they did show the cumulative percent of patients in whom fever had resolved, and my issue with the paper, I like the fact that at two, three, four days cumulative percent fever is resolved.

It seems like a reasonable endpoint. They didn't really describe what the definition was of fever being gone, so I don't know below which cutoff point that criteria was -- what criteria was utilized.

And then this is the more traditional endpoints, especially in hospitalized patients, returning to stability or normality of things like the oxygen

saturation, the heart rate, blood pressure, respiratory rate, and that part is obviously for maybe tomorrow's discussion in terms of severe pneumonia, and these are the outpatients, and the question is, would we be able to capture that data in outpatients.

How reliable would the patient population be? I don't think folks who are involved in clinical trials would want nurses going to patients' homes three or four times a day to take their temperature, so how reliable would the measurements being conducted by the patient be?

We also want endpoints that document a clinical failure and/or drug adverse effects, and we're going to hear after lunch about adverse effects, so I won't dwell on that, but we also need objective evidence of failure.

We've already heard about regression of the infectious process manifest clinically by patient-reported observations

that don't improve or obvious evidence of microbial invasion, and I just put up as examples empyema, bacteremia, meningitis.

The micro biologic endpoints we've already said are not of value as far as success or cure, but they certainly become important at documenting clinical failure of therapy.

So useful success endpoints for mild to moderate community-acquired pneumonia trials. Patient-reported observations are a valid, reproducible, and meaningful outcome measurement tool that seems to deserve increased utilization. If carefully implemented to ensure reliability, time to resolution of fever and pertinent laboratory results seems reasonable, and we heard some examples of that earlier today.

Things that don't seem useful as successful endpoints are mortality, the radiographic response. Microbiologic response for cure may be valuable for failure and

"return to usual activities," since that is so highly variable.

Failure endpoints. Trial design should be able to detect failure of therapy, as well as success and have the ability to detect adverse events, microbiologic data, again, documenting clinical failure.

So it's crucial that valid clinical endpoints support claims of efficacy of new anti-infectives. The use of patient-reported observation should improve endpoint data.

Improvements in rapid specific identification of the microbial etiology of community-acquired pneumonia will increase the likelihood that observed clinical responses represent a treatment effect. Thank you.

DR. COX: Thanks, Dave, and we'll take questions for Dr. Gilbert, and maybe I'll start. I think maybe you had a question for me in your presentation, and I'll try and respond to it a little bit.

DR. GILBERT: Please.

DR. COX: Then we'll go to Dr. Powers. Your question was about the use of PRO instruments and NDAs that have been approved based on that, and certainly in other therapeutic areas PRO instruments have been used.

In the anti-infective area, I mean, a lot has relied on physician global assessments. The physician may ask the patient questions on a variety of different symptoms and make some assessment of how the patient is responding, but in general PRO instruments have not been something that we've seen much of.

DR. GILBERT: So it's true that, to date, you have not looked at a new drug application that uses PROs as part of the database?

DR. COX: Well, a PRO instrument.

DR. GILBERT: How do you evaluate colds and stuff.

DR. COX: Well, that's -- well, for

influenza, that's true, yes. In the label for influenza there is a --

DR. GILBERT: See, those virologists have all the advantages.

DR. COX: Yes, sorry about that. It's hard to index all these drugs, but although formal PRO instruments have not been used, people are looking at patient symptoms.

It's just that, typically, it's the physician asking the patient what their symptoms are like, and then the physician records their impression of what the patient is reporting to them, so there's the intermediary in their impression.

Dr. Powers?

DR. POWERS: Dave, I wanted to address the issue that you brought up of are there other patient-reported outcome instruments in community-acquired pneumonia. There are, but the CAP-Sym one that you presented has one huge strength that the other ones don't, and that is that they actually

interviewed patients in focus groups to get their impression of what they thought was important, as well.

So when you want to put one of these together, it really entails doing three things, getting the patient's point of view, getting the clinician's point of view, and doing a literature search on what are the appropriate elements to include, and that one actually does.

The other thing is there is actually a scaled-down version of this that only asks 12 questions, too, so it would be really nice to look at whether the scaled-down version operates well, too, so we could ask people fewer questions, but there is one thing it doesn't do, and I wanted to ask Bob and Bob to comment on this. It's how much of a change in that instrument is clinically meaningful, and that's the one thing that's missing.

I know Bob and I did something on this before at DIA about evaluating the entire

distribution versus responder criteria, and so when you're developing one of these for use in a Phase 3 trial, it's very helpful to test drive it earlier on in earlier phases to try to figure out how you're going to analyze the results.

So what was presented in the Torres trial is, okay, it goes down from 34 to 20. Well, what does that mean for people? And that's the one piece of information that we're lacking here.

DR. COX: And I'll ask Dr. Temple to respond.

DR. TEMPLE: Well, our guidance urges people to learn what the minimum important difference is, and you do that among other ways by asking the same patients that you developed the scale with. How much difference seems to matter to you? And that's what you do.

Just, by the way, there are lots of places. There is a scale for heart failure

symptoms called Living With Heart Failures developed by NIH. That appears in the labeling. Arthritis drugs typically use them. I mean, any pain scale is a patient-reported outcome. Who else knows about the pain?

So they're coming all over the place, and, as John suggests, one of the concerns is they're so sensitive maybe you can pick up differences that are utterly trivial, although finding anything is something, you know, in some ways.

DR. O'NEILL: On that, I think, John's right, probably. One of the things that struck me, and I would -- not knowing how they validated this instrument, I was curious in this particular situation; if you had a series of focus groups or test groups or whatever, one of which was -- really had CAP and one of which really didn't. You knew the true state of nature, and you asked them the same questions, and you validated this, this instrument.

It doesn't weed out who has got it and who doesn't have it. It answers another question. It sort of says, "Symptomatically, do you respond to treatment in some sense?" So it would be interesting to see the disconnect between the patient response where, if you have a mixture -- and that goes back to my original question that I was talking about earlier.

The problem here is that these trials enter two types of patients, those that do and those that don't have the infection, but you count them all up at the end of the day, and your trial is going to rise or fall on what percentage mixture you have, and if you want to go to the throw-out game, which isn't such a great idea, and that's what I was getting at, the sensitivity and specificity of the classification, but where I'm coming at is if the PRO is applied to the mixture population, would the folks who do and do not have actually CAP in the population respond in

a similar way?

And you could actually do that at the validation stage, so I was curious whether they did that, and following on what Bob Temple said is you can ask these people the same thing, is what is the minimal difference where they can discriminate benefit versus non-benefit.

So how do you go about using this, this focus group, to get the distribution among that crowd as to what you would feel personally that would be a meaningful benefit from when you began and when you ended? And I think that part of that philosophy is in the guidance, so to say, "You need to be doing that up front," and it may not be one number.

It's probably a distribution that you need to get a sense of a representative population in terms of what people value as a meaningful change in whatever the scoring mechanism is.

DR. GILBERT: In the Lamping study,

and John Powers will correct me if I'm wrong, they did not validate the instrument in well people. They only looked at sick people, but John participated in this symposium at the Mayo Clinic where they did address such issues, and I don't know if they focused in on community-acquired pneumonia or not, but they did talk about the need for the very point that you're describing, that well people versus sick people should be part of the validation.

DR. O'NEILL: They all feel sick. They're not really well. They all feel sick. It's just that the clinical diagnosis is such that you enter them in, but do they or don't they have the bug, and that's what I was trying to get at, a sense of whether those two --

DR. GILBERT: But I thought your question --

DR. O'NEILL: -- classes of folks respond differently.

DR. GILBERT: I thought your question was what are the subjective symptoms that pneumonia patients consider important, and so you would want to have a panel of well people to ask them what should the symptoms be in our questionnaire instrument. I thought that was your -- part of your question.

DR. COX: Dr. Fine?

DR. FINE: So, I have two questions and a comment. One is you mentioned a lot of patient report surveys that had to do with pneumonia-specific symptoms, but there is a whole literature out there on quality of life instruments such as the Medical Outcome Study Short Form 36 and Short Form 12 that allow us to have population-based estimates of quality of life, and do you see any role for those in assessing outcomes?

And then a comment is that one way, in addition to focus groups, to validate the importance of a given delta in symptom reporting would be to see how they correlate

with population norms and things like these generic quality of life instruments. That would be one approach.

The other approach would be to use standard decision analytic techniques where you do time tradeoffs or willingness to pay or other forms of utility assessment to actually get a sense from patients how much a change of ten points in a symptom scale, how much meaning that actually has.

DR. GILBERT: I didn't delve into the quality of life instrument. Obviously, it's not as focused on pneumonia, and so I didn't think it was as relevant. With the time constraints, I didn't delve into that.

So I know that in the Torres study that it used this validated instrument. They did compare the quality of life instrument with their questionnaire, and they got smaller standard deviations with their questionnaire than with the standard quality of life questionnaire.

I think Rich is next, actually.

DR. WUNDERINK: Rich Wunderink from Northwestern. I just want to make a comment. I think that this is probably valuable, but it goes back to one of the things that Tom said earlier about using the PCT as an endpoint, and that is, antibiotics have other effects besides whether they're killing the bugs, and you may have both a beneficial effect, say, with some of the anti-inflammatory properties of a macrolide or an adverse effect, say, some of the dysphoria that quinolone patients may experience that will adversely affect whether this score really is determining are they responding to an antibiotic.

Now, the point of all of this is making patients feel better, and, you know, the issue may be we trade pulmonary side effects from the pneumonia and increasing drug effects, and that's why I'd echo what John said, that we probably need to have these done

on a regular basis, at least daily if not a couple of times a day, because you may actually see a shift in the symptoms that actually get -- points as they make that transition, and so it's going to be a little bit more complex than just a composite score.

DR. GILBERT: Well, it's interesting that you brought up the quinolone confusion point. I didn't show all 18 questions, but about four of them were mental status questions, so in your -- and confusion is on that scale.

So in your example, the patient would be getting better on 12 of the symptoms that are listed on the questionnaire, but the confusion questions would be scored. The confusion score would be staying higher, even increasing, so it still might work.

DR. COX: Okay, Dr. Bradley?

DR. BRADLEY: John Bradley from Children's Hospital in San Diego. I am the pediatrician up here on the panel, and coming

into these discussions, having spent four years on the FDA's Anti-infective Advisory Committee, I'm acutely aware of how important the FDA's guidances are in drug development, and as it was mentioned earlier, it's the clinical scenarios that we see that drive the need for drugs.

Clinical science, the trial designs need to be scientifically valid so that we can see which drugs work, but at the end of the day, all of these complexities that we're talking about, and clearly within adults there is complexities that the PORT scores have documented in pediatrics.

A six-month-old is different than a two-year-old is different than a five-year-old, and in order to do all of these clinical trials for all of these subgroups -- and we knew this was complicated coming into this meeting, and the complexities are just exponentially increasing.

At the end of the day, it's Ed's

job to come out with a guidance for everybody for community-acquired pneumonia, and that's a virtually impossible task, but the scientific design absolutely has to be impeccable.

Dr. O'Neill was saying, you know, he doesn't want non-inferiority. He wants superiority trial designs, but the points that Dr. Fleming made about value to drugs that may not be all in getting you a patient cure with no fever at four days are very important, and in pediatrics, taste of an antibiotic that's given by mouth is a huge issue. A horrible drug which is poorly tolerated is of no use to us at all.

So all of these complexities clearly come back and reflect on the clinical trial design and the regulatory design. The regulatory design impacts drug development, and the drug companies in here are going to be looking at the regulatory issues before they actually jump in and develop new drugs.

So I am here to try and say that as

I'm listening to all of this, it's the compromise that we're not going to get the perfect scientific studies, but we need the drugs that I'm going to be enrolling patients in studies, and for a dose-ranging study where there is going to be failures, I have to talk a mother into saying, "We're starting out with the low dose. We're doing this so some of these children who I am treating can fail. Please sign on the dotted line so I can treat your child."

That's very difficult for me, placebo-controlled study, yet I know we need to know what the placebo -- what the benefit of the drug is in order to correctly evaluate whether the toxicity of the drug is balanced with the clinical benefit.

So what I'm doing is expressing a need for compromise, not that each of these points isn't critical, but at the end of the day, we're going to all have to come together to something that's acceptable is not perfect.

DR. GILBERT: John, I'm glad you brought up "the end of the day." So at the end of the day today, we are going to have a panel discussion, and we're going to start at this end with Dr. Niederman and go around the table and give you each three minutes to answer the questions that are on the end of today's program so that all such sentiments as you so eloquently just expressed will come out. So be forewarned; you will be called upon. Dr. Psaty.

DR. PSATY: Bruce Psaty from Seattle. I don't normally read antibiotic trials, but I read a large number of them in preparation for this meeting, and the standard outcome is an investigator-determined resolution of symptoms, and it would be actually very useful to know how some of these outcomes map to those, and what look like the thresholds for the determination of cure. That would be another way to look at it.

It would also help to move the

trials back so we could have indirect comparisons with the previous trials. If we're talking about doing non-inferiority designs, and we have to map backwards in time, it would be very helpful to know how to do that.

The additional difficulty is, it's a continuous outcome, rather than a failure time model, so the statistics are a bit different, and you can have missing data on these, and some of it can be informative missing. What do you do when someone dies, and how do you handle these scales in the face of death or missing data?

DR. GILBERT: Actually, many of those issues are in that supplement of Value in Health that I made reference to, how to deal with missing data, how to deal with the patient that has an unexpected adverse effect and so forth.

I think Ed has a very important lunch announcement.

DR. COX: Yes, I was going to do one last question here from Dr. Fleming, and then we'll get to the lunch announcement, because we do want to be on time, so one last question from Dr. Fleming.

DR. FLEMING: Okay, it's as much a comment as a question, which I'm partly reluctant to give, because I'm actually a great fan of PROs, and this sounds more like bringing out the concerns, but there has been a great interest in PROs across disease areas.

The oncology area has been studying this intensively for a long time, and essentially what approvals are based on now would be survival and disease-related symptoms like pain or, in prostate cancer, skeletal-related events like fractures, pain, or spinal compression, and the challenge in being able to get the richness of other elements into the PROs has been a common experience seen across all disease areas.

First of all, you have to have a

blinded trial, and you have to be assured that the patients, when they're giving their assessments, the blind, the integrity of the blind is maintained. Secondly, there is inherently a lot of difficulty in avoiding missing data, and the missing data here is informative missingness.

Thirdly, it's a multiplicity of components, and it's difficult to statistically understand multiplicity, so what do we do? We look at a composite. Well, as soon as you look at a composite, interpretability becomes a lot harder.

And so, for example, with ximelagatran in knee replacement, they had a composite endpoint that looked wonderful, and when you looked at it, 90 percent of the events showed a great effect on asymptomatic distal DVT, and nobody knew what that meant, but when you looked at the important events, death, stroke, MI, major bleeds, pulmonary embolism, it went in the wrong direction.

So when you have many different components in a composite, if they're apples and oranges in terms of what their importance is, it's difficult to say what's a clinically meaningful change, because if the clinically meaningful change is based on something like sputum versus something like what would be breathlessness or something that's much more important to the patient, that's going to make it harder to define what's clinically meaningful, so when we do these composites, you need likes with likes in terms of comparable importance in order to be able to interpret this.

So I hate to end with the concerns, because I'm truly enchanted with the idea of trying to advance PROs. Why? Because they are specific to what it is that the patient really experiences, and that's what we should be looking at.

DR. GILBERT: But can't you group the PRO items so you can --

DR. FLEMING: You can.

DR. GILBERT: You can group them --

DR. FLEMING: You can.

DR. GILBERT: -- into the highly specific symptomotology, the pneumonia, and then we'll have a group for adverse effects, et cetera.

DR. FLEMING: Yes.

DR. TEMPLE: One of the pieces of experience that we've had is that patient-reported outcomes that are developed for a particular disease work a lot better than some of the quality of life things, the SF36 or whatever it's up to now, and they've been very successful in asthma, where they are very well targeted and developed, this Living with Heart Failure. There are a number in arthritis that are similar, and those are very good.

The general quality of life things really don't work out, because most of the -- most of your psychiatric state and your ability to interact with your neighbors just

isn't affected by this stuff, so the targeted ones work much better.

DR. COX: Thank you. So we're at 12:15, and just to let folks know, I've heard it's snowing outside, so there is lunch that's available in the lobby level in the River City Grille, and it sounds like what they're going to have set up is going to be a lunch buffet at $12.95, so that may be one of the more convenient options for lunch today.

There are other restaurants in the area, but it's a short walk, and if it -- I haven't looked out there to see what the snow is like. The option downstairs might be the best way to keep dry and warm.

DR. GILBERT: But the haunting reality is 45 minutes; right?

MR. COX: Yes, it's only 45 minutes. We'll be back starting promptly at 1:00.

(Whereupon, the foregoing matter went off the record at 12:20 p.m.)

DR. GILBERT: In order to give Dr.

Murphy his full measure of time, I think we really must get started.

Tim is kind enough to travel here from Buffalo, where he's a, I believe, distinguished Professor of Medicine and Microbiology and Chief of Infectious Diseases. Tim, thank you for joining us.

DR. MURPHY: Thanks, Dave. When I looked out the window and saw the snow, it made me feel like I was at home, except we measure our snow in feet.

All right. So, my mission here in the next 25 minutes or so is to answer the question on the slide there, "Does the literature document a treatment effect relative to placebo in community acquired pneumonia?"

So, it's a nice opportunity to do a review of the literature, the old literature, the new literature and give you a comprehensive and scholarly answer to that question, which is no, okay. There are no

placebo controlled trials of community acquired pneumonia.

So, I rephrased the question in the spirit of this meeting and discussion. So, I think what we would like to know is, are antibiotics effective in community acquired pneumonia and then, in terms of thinking about whether placebo groups are rational and reasonable, what is the etiology of mild to moderate community acquired pneumonia and specifically, what are the relative roles of the so-called typical bacteria, atypical bacteria and viruses.

Then finally, I'll address the question, should placebo controlled trials be performed in mild to moderate community acquired pneumonia?

So, are antibiotics effective in pneumococcal pneumonia? This is a graph actually, that's reproduced in modern software, I suppose, from the classic study of Austrian and Gold in the Annals of Internal

Medicine in 1964.

So, the green bars are treated with penicillin and then the gold bars come from another classic study from 1937 in the pre-antibiotic area, and you see by sero-type and you see all types, but penicillin is dramatically effective in bacteremic pneumococcal pneumonia. That's not necessarily mild community acquired pneumonia, but I'm going to continue to address that point.

So, penicillin is effective for pneumococcal pneumonia and Dr. Austrian and Gold made an interesting statement in the discussion, which is pertinent to our discussion here, which is, it is questionable that a more effective anti-pneumococcal drug than penicillin can be developed.

In 1964, there was no penicillin resistance, but their point really was that very low concentrations of penicillin are highly bactericidal for the pneumococcus and

from that standpoint, they speculated, we're not going to do much better.

So, what is the etiology of mild to moderate community acquired pneumonia and what are the relative roles, as we've used this morning? Typical bacteria are generally Streptococcus pneumoniae and Haemophilus influenzae and the atypicals are Chlamydia pneumoniae, Mycoplasma pneumoniae and Legionella is an atypical, but really behaves more like pneumococcal and Haemophilus influenzae pneumonia.

So, if you look at the slide here, you'll see the usual diagnostic criteria in many, many studies. So, I looked at it, as we probably all did, a lot of these comparative trials, that have been published in community acquired pneumonia and for so-called typical bacteria, a positive blood culture, which is, in mild pneumonia, is a very low sensitivity, maybe one percent positive, and then a positive sputum culture of an adequate sputum

sample, which is usually described as greater than 25 neutrophils per high powered field and less than 10 squamous cells in a Gram stain. That's presumptive diagnosis.

Atypical bacteria, the usual way in many, many studies is to look at a four-fold rise in anti-body titer, so this is Chlamydia and Mycoplasma, or an elevated single level in a single sample.

The problem with these, as I'm going to point out, is that if you look at the typical bacteria, we're only looking at less than half of patients for a possible etiology, whereas for the atypicals, we're looking at almost everyone, because you can get blood samples.

So, let me -- so, what I did then is looked at all of those -- or many of those, my big pile of mild pneumonia comparative trials, to try and make -- I was going to make a big table to show how many sputum samples were assessed in each one of those studies,

and I'll tell you, you can't tell.

In the vast majority of these studies, you absolutely cannot tell. They'll tell you how many pathogens are isolated, but almost never, how many sputum samples were assessed.

I found two studies, and these are actually good studies, relatively speaking, in terms of trying to reach a diagnosis. The one on the left is Falguera and colleagues in Archives of Internal Medicine and this one is Rosen and colleagues in CID.

So, they used sort of -- this is their criteria for a positive sputum blood or pleural fluid culture in this study, and it turns out that 27 percent of their patients produced a sputum sample that meet the criteria and nine percent of pleural fluid. So, there is a lot of overlap there, but let's say, a third of patients then, were evaluated for pneumococcal or H-flu pneumonia, and this study actually did blood culture, sputum,

pleural fluid, and they did some transthoracic cultures, and you'll see, with lots of overlap, perhaps half of their patients were evaluated for typical pathogens.

In the atypical, they both used serologic. This one also used a PCR of a throat swab. So, both of them evaluated about 80 percent or so of patients for atypical pathogens.

So, if every culture here is positive for the pneumococcus, then one-third of these patients would have pneumococcal pneumonia. Then add in the antibiotic, the antibiotic that was given before.

So, in this study, they didn't tell us how many people got antibiotics, but it was not an exclusion criterion. So, you could knock this down by an unknown percentage, because we know probably one dose of antibiotic will render a sputum culture negative for pneumococcus or H-flu. Maybe not one dose, but it doesn't take much. These

guys told us 27 percent received antibiotics. So, we can drop this 50 percent down another 27 percent.

The bottom line is, we are very much -- well, before I say we're under-estimated, let me just say, we are not adequately testing people for pneumococcal and H-flu and Moraxella pneumonia in our studies of community acquired pneumonia. So, the usual diagnostic approach in studies of CAP under-estimate the proportion of typical bacteria.

So, let me show you three studies that went the next step, very nice studies that looked harder, shall we say, for typical bacteria.

One was by Gutierrez and colleagues in CID in 2003 and they looked at 493 patients who had community acquired pneumonia by good criteria, new infiltrate on the chest x-ray, and they attempted to determine the etiology in pneumonia and they really looked at the

pneumococcal urinary antigen and in particular, studied pneumonia of unknown etiology.

So, they used the usual criteria that I just described to you. Here is the pneumonia severity index. Three-quarters were PSI one, two and three. They identified an etiology in 40 percent. That's probably average, based on these kinds of studies, and they studied urinary antigen for pneumococcus, in particular, in the ones -- well, in everybody, actually.

So, you see the results here and what I'll tell you then, without going through this in detail, they calculated a 70 percent sensitivity and a 90 percent specificity for urinary pneumococcal antigen. In fact, if you look, Pseudomonas and other Gram negatives, I would question whether those are really etiologic. They probably have a higher specificity than 90 percent.

So, that's their sensitivity and

specificity, and they looked at the 300 patients that had pneumonia of unknown etiology and 23 percent of them were positive in urinary pneumococcal antigen.

So, a proportion of patients who have pneumonia of unknown etiology, that pneumonia is caused by the pneumococcus, would be the conclusion that I would reach.

Second study by Ruiz-Gonzalez and colleagues in the American Journal of Medicine, they did microbiology a study of lung aspirates, of trans-thoracic aspirates in 109 consecutive patients with community acquired pneumonia over a 15 month period.

They used serology to make a diagnosis with atypical pathogens. Their patients, we don't know pneumonia severity indexes, but the mean age was 51. They said 44 percent have underlying illnesses, meaning what is it, 56 percent don't. Twenty-nine percent were treated as outpatients, but we don't know the criteria they used and the

group had a low mortality. Very interestingly, 43 percent received antibiotics before the procedure.

So, on their assays of these trans-thoracic aspirates, which they did on 109 consecutive patients, except people who had contraindications and those who refused, they did a bacterial culture, a selective culture for Legionella, they did capsular antigen detection, for the pneumococcus and for H-flu Type B, which is really not a very common cause of community acquired pneumonia, it's mostly non-encapsulated and non-typable H-flu, though they found a handful, and they did PCR on the pneumococcus and they did PCR for these three atypical agents, and here's what they found.

So, actually they looked and out of their 109 patients, by conventional testing, they made a diagnosis in 54, Mycoplasma pneumoniae, Chlamydia pneumonia, pneumococcus, and with the trans-thoracic aspirates then,

the distribution changes dramatically. Pneumococcus is the most common. They made a diagnosis in 90 of 109, followed by Mycoplasma, Chlamydia and then Haemophilus influenzae enters into the top five pathogens.

So, 33 percent of the patients without an etiological diagnosis by conventional means had pneumococcal infection detected by one of those methods, in a trans-thoracic aspirate and I would argue, it still underestimates the proportion of typical pathogens.

Forty-three percent received antibiotics. The PCR probably was not too much affected by antibiotics, although it may have been. PCR was only done for the pneumococcus, not for H-flu, not for Moraxella and not for other pathogens, and antigen detection was only done for Type-B H-flu, which is not a particularly common cause of community acquired pneumonia.

Then the third study that I'd like

to show you is perhaps the best in a way, in terms of taking a good look for typical pathogens and this is Lim and colleagues, published in Thorax in 2001. So, they made a big effort to look for bacteria.

Positive blood culture, pleural fluid culture, positive sputum culture. They did counter-immuno-electrophoresis for Streptococcus pneumoniae on sputum samples, looking for capsular polysaccharide. They did serology on the pneumococcus. They looked for a three-fold rise in antibody titer to three pneumococcal antigens, C-polysaccharide, pneumolysin and pneumococcal surface protein A, PSAA, and they looked for a three-fold rise in antibody titer to H-flu and Moraxella catarrhalis, using a laboratory strain, which I applaud their effort, but a lot of these immune responses are strain specific. So, unless you use the patient's own strain, you're going to miss a lot of immune responses. But they actually did find a

couple in that group. The used the usual criteria for atypical bacteria.

So, what did they find? They found, in fact, that typical, this is -- I'll show you the bugs in a minute, typical is twice as much as the atypicals, when you look hard for pneumococcus in particular and Haemophilus influenzae, followed by atypical, followed by viral, followed by no pathogen.

These are the bacteria, far and away, Streptococcus pneumonia, Haemophilus influenzae, Moraxella catarrhalis and then they found some Staph aureus and Gram negatives.

So, that's three studies, actually, that look harder for typical bacteria, by pneumococcal urinary antigen, by trans-thoracic aspirates with PCR and then, by serology and antigen detection and each one of them shows a substantially larger proportion of people with typical bacteria pneumonia, in particular, pneumococcal pneumonia than is

apparent when we look at studies of community acquired pneumonia.

This last study, the Lim Study, very interesting also, looked at the value of the diagnostic test. So, I'll make one point here, and that's -- they looked at it with prior antibiotics and no prior antibiotics.

So, three patients' blood culture, look at the sole mean of diagnosis, urinary antigen. The asterisks means, interestingly, that they had more samples positive, statistically significant in prior antibiotic -- no prior antibiotic, compared to antibiotics, urinary antigen, in particular. The only one that was unaffected by antibiotic was the serology.

This was a penicillin resistant pneumococcus. So, we know that prior antibiotics are going to really have a dramatic effect on cultures and counter-immuno-electrophoresis, interestingly, probably by reducing the titer of bacteria was

also different in the antibiotic versus no antibiotic group.

Again, many of the trials that we see, that look for diagnostic -- look at -- do diagnostic studies, people have received prior antibiotics. It's important to keep this in mind.

So, what about viruses? I showed those three studies with typical bacteria. What about virus as a cause of community acquired pneumonia?

We heard Dr. Nolte's excellent discussion this morning about the molecular diagnostics for viruses and that is the future, in terms of sorting out viral etiology of respiratory tract infections.

The problem we have now in 2008 is the interpretation of the results, and let me point out a number of limitations where I don't think we can make an intelligent statement about how much viruses are causing community acquired pneumonia.

So, when you look in studies who look for viruses, it's generally a sample recovered at the time of the pneumonia, either sometimes a sputum sample, usually it's a nasopharyngeal swab or a throat swab, single result.

So, if we look at trials of -- studies of COPD, where you do people when they're sick and when they're well, several studies from several groups show that you can find viral RNA in up to 15 percent of clinically stable patients. These are good studies with good controls. The viral RNA is there. That's not the question. The problem is, the virus is probably not doing anything. It's not making the patient sick.

Interesting study from Dr. Macek and Jim Hogg and colleagues, in the Canadian Respiratory Journal in 1999. They looked at 20 lungs post-mortem, people who died of asthma and actually, some of other causes, and did PCR with good controls, looking for nine

common respiratory viruses. Nineteen of these 20 lungs had viral RNA. Fourteen to 20 of them had two or more viruses in the lungs. Point being that their speculation is, the lungs are a reservoir for common viruses.

In bacteria, we would describe this as colonization. I mean, the viruses are there, but they're not causing disease. So, we need to be very careful about how we interpret positive viral RNA in respiratory tract samples in people with community acquired pneumonia.

I told you the first two. We know in studies of COPD, the people sero-convert and they have asymptomatic viral infections all the time. Most of the sampling is done on nasopharyngeal and throat samples, and so, we know viruses cause upper respiratory infection. We don't know how often they get into lower respiratory tracts. So, I think we need to be critical, in terms of evaluating the samples that are being studied.

Then, we all speculate and think and say, though the data are hard to come by, that preceding viral infections pre-disposed to bacterial pneumonia, I think that's probably very likely true, but when you look for actual hard data, that's tough to come by.

So, a viral PCR on an upper airway sample is not going to be able to make that distinction as well.

So, my conclusion actually is that currently, there is little convincing evidence that viruses cause a substantial proportion of community acquired pneumonia in adults. I was careful in my wording. I didn't say it's not causing it, but I don't think there's very good evidence at this point right now.

So, what is the etiology of mild to moderate community acquired pneumonia and what are the relative roles of typical and atypical bacteria? Most studies underestimate the proportion of typical bacteria because of limitations in diagnostic studies, as I

pointed out and I would conclude that bacteria are the predominant cause of mild to moderate community acquired pneumonia.

In particular, pneumococcus and Haemophilus influenzae, I would estimate, cause well over half, perhaps up to 75 percent of the community acquired pneumonia, based on the data that I showed you.

So, should we be performing placebo controlled trials in mild to moderate community acquired pneumonia? I think that was sort of the question that I was asked in so many words.

Let me make a point here and it's -- in our discussions this morning even, sometimes we are lumping together exacerbations of COPD with community acquired pneumonia. Those are very different diseases with a different pathogenesis and different distributions of pathogens.

The third most common cause of -- bacterial cause of exacerbation in COPD is

pneumococcus. Viruses undoubtedly play an important role in exacerbations of COPD.

So, this community acquired pneumonia is a very separate disease from acute exacerbations of COPD. I've actually worked hard to try and facilitate performing placebo controlled trials in exacerbations of COPD.

I'm going to make the argument that we should not be performing placebo controlled trials for community acquired pneumonia, mild, moderate, severe, any severity. One, because the predominant cause is a pneumococcus, and we have effective therapy for the pneumococcus.

Number two, there is the potential for adverse outcomes. Even look at the pneumonia severity indexes, there is a small mortality associated with mild community acquired pneumonia and all of us who take care of patients with pneumonia see the occasional patient that doesn't follow the rules. Some

of them get sick. So, it's difficult to withhold effective therapy, in a disease where the estimate is 50 to 75 percent are caused by a treatable bacterium.

As the discussions have revealed this morning, faster recovery and return to baseline are clinically important outcomes. Mortality is not a meaningful endpoint, in mild to moderate community acquired pneumonia. I think the patient reported outcomes are going to be the way to go, rigorously done.

If a person gets better and then goes back to work in two weeks, compared to one week with antibiotic, that's a significant endpoint. That's not a good endpoint. It's with the more rigorous endpoints we're talking about, in patient reported outcomes, and from a pragmatic and a practical standpoint, the reality is, many physicians and many investigators would balk at placebo controlled trials for community acquired pneumonia. It will be difficult to enroll patients in such

trials.

So, my conclusion/opinion is, where we are right now in 2008, with our ability -- or shall we say, our inability to make an etiologic diagnosis of community acquired pneumonia, my view is that we should not include a placebo group in community acquired pneumonia for any severity of pneumonia. Thank you.

DR. GILBERT: Thank you, Tim. One quick question, the direct lung stick studies -- and I'm thinking Spain and Japan, I haven't reviewed that literature recently, but they also found, to me, a striking percentage of rhinovirus, and you think that's just colonization?

DR. MURPHY: Yes, it's hard to know. There's a very interesting and evolving, a nice rhinovirus literature that -- it's sort of ironic.

In COPD, I think we know more about what rhinovirus is doing, compared to in

community acquired pneumonia.

Rhinovirus, we think of as an upper respiratory pathogen. In COPD, it definitely gets into the lower airways and enters respiratory epithelial cells.

In community acquired pneumonia, I don't think we know that yet, and so, I think the jury is out, actually, in terms of understanding what that means. Though when you get it out of lung aspirates, I think you have to pay attention to it. That's different from a nasopharyngeal or a throat swab.

DR. GILBERT: We have time for maybe one question. Yes, Rick?

DR. NOLTE: In the COPD study, you said the viral normal pleural was what? Which viruses?

DR. MURPHY: So, the whole -- so, it was rhinovirus, para-influenzae virus. There is actually a fair bit of RSV in some studies and there is some controversy about that. Some people think that -- some folks with COPD

are chronically infected with RSV. Other groups don't find that. Coronavirus is the other one, the main viruses.

DR. NOLTE: The other thing that's really becoming -- I think it's going to be interesting, as these new tools are developed to allow us to cast that wider net is opportunity to see mixed infections grow sort of, exponentially, not only mixed viral infections, but the contribution that perhaps, the bacteria and the virus make together, in terms of the presentation of the disease.

DR. MURPHY: I absolutely agree. I think that's one of the areas of great fruitful investigation, is looking at the interaction of viruses and the bacteria and I think that some bacteria cause infection only when there's a preceding virus and there is probably vice-versa, particularly in COPD, people who are chronically colonized in the airways by bacteria, there are a lot of proposed mechanisms for why those folks are

more susceptible to viral infections.

So, I think looking at the interaction of viruses and bacteria in the pathogenesis of respiratory tract infection is a key area to study.

DR. GILBERT: Roger is going to ask you a question and he promises to be brief.

DR. ECHOLS: Tim, based on your review, would you try to characterize the different severities of pneumonia by different etiologies or that they are of similar etiology?

DR. MURPHY: Yes, good question. I looked at that, and you know, it's one of the questions that I wanted to address. There's not a good answer to it, based on firm data, but it does look like certainly, that younger people have more atypical pathogens, compared to older people, and atypical, Mycoplasma and Chlamydia, probably may cause a less -- likely cause a less severe pneumonia than the typical bacteria, okay.

But in several studies, when they looked -- in fact, one of them that I showed the distribution of pneumococcus among all the PSI types was identical. Other studies show there is less pneumococcus, but those studies are limited by the things that I said, mainly because we're only looking at maybe a third of them for even the presence of the pneumococcus.

DR. COX: Dr. Murphy, one more question if I might. It's a difficult one and it gets to the issue of, if placebo controlled trials are not really something we can consider for patients, regardless of severity, what would -- do you have any insights or suggestions, what the control group might be or what the design might be in those populations? It's something we've struggled with and I'm just curious what your thoughts are.

DR. MURPHY: It's a tough one and -- sure, and the non-inferiority margin as well,

sure. I mean, again, it's out of my area of expertise, but I'm in a room full of experts and it would seem to me that a well designed, rigorously performed, non-inferiority trials with the standard of therapy, looking at patient reported outcomes would be a rational way to approach antibiotic trials for new antibiotics in community acquired pneumonia.

There are lots of different potential nuances, okay, and the discussion has been also, so, could we eliminate people with PSI-1, for example, or eliminate the mild pneumonia and just look at a sub-set of people with community acquired pneumonia?

Again, that would be my simple-minded answer. There are a lot of limitations to it, clearly.

DR. GILBERT: Two people have promised to be short. All right, Tom, short.

DR. FLEMING: My understanding of the basis for this conclusion -- although I wish we had a lot more time to probe on that,

was that we -- basically, we know that we can treat pneumococcus. That actually doesn't mean that we can't do a superiority trial.

If you think that we would be using a standard of care that would be effective for pneumococcus, you'd still be able to do a -- I'll call that agent-A, A plus B, against A, which technically is a type of placebo controlled trial. It's an add-on trial.

So, I didn't interpret this to mean, you had to do non-inferiority. I interpreted this to mean, you believe that there are effective agents for pneumococcus, and we don't want to deprive patients of access to those in the trial.

DR. MURPHY: Agreed, absolutely. I think it would be possible to do superiority trials with patient reported outcomes.

DR. GILBERT: Dr. Rex.

DR. REX: My question is just for clarification. Your survey, you concluded there were no data for CAP, for placebo

controlled, for any level of severity or just for mild/moderate?

DR. MURPHY: I believe for any level of severity. I didn't find any placebo controlled trials for community acquired pneumonia.

DR. REX: That's what I wanted to hear you say. There are no placebo controlled trials that you can find, that are meaningful for community acquired pneumonia?

DR. MURPHY: Correct.

DR. REX: That's the bottom line.

DR. MURPHY: Maybe others have found it.

DR. MUSHER: Could I add something to that? I don't mean to be argumentative, but in the J. Burns Amberson Lecture that Dr. Finland gave in 1979, he did comment on the placebo controlled trials and the amazing thing is, he showed very little difference in the outcome, the placebo controlled versus penicillin treated, which blew me away.

I absolutely agree with you. I don't think -- you had there -- the physicians are reluctant. I, as a patient, would be damned if I'd have signed the consent form for a placebo controlled trial.

I do want to point out that the Swedes recommend for mild to moderate community acquired pneumonia, the recommendation is penicillin, to support Dr. Murphy's point.

DR. GILBERT: Okay, we must move on. Ed?

DR. COX: Okay, great, and John, we'll come back to you for the first question afterwards. Next speaker is Karen Higgins. She's from FDA. She's a Statistical Team Leader in the Division of Special Pathogen and Transplant Products, and Karen will be talking to us about statistical issues and endpoint selection and non-inferiority trial design from an FDA perspective. Karen.

DR. HIGGINS: Hi. The title of my

talk has changed a little bit from the agenda, but it's the "Overview of Recent CAP Trials, Non-inferiority Trial Design and Endpoints." I was told that you'd be fairly interested in getting a summary of what we've seen at the FDA for non-inferiority trials for CAP.

My outline, I'm going to go over some of the issues with non-inferiority trials that we can -- that we think about at the FDA. I'll go over it briefly, since Dr. Fleming gave such a nice presentation, then review recent adult CAP trials and the bulk of my talk will be on oral-only studies. Of course, they would be the more mild to moderate CAP. I'll go over their study design and the results that we saw. I'll then briefly review the IV to oral studies and end with a summary.

Note that of my review, I've attempted to mask the studies, so that none could really be identified and I collected this information from FDA reviews, study reports, data submitted and data collection

was not the same from NDA to NDA. So, some information was not readily available or available at all. So, therefore, please take it as an overall summary of these studies.

So, the goal of a non-inferiority trial is to show the efficacy of a new drug or a test drug and we do this by showing that the new drug is similar enough or in general, not too much worse than the control, in a well-designed and conducted trial.

And so, what's needed to do this? Two very important things. The first is to have information on the efficacy of the control drug. This is based on historical information, past placebo controlled studies, preferably, and this is what's needed to justify a non-inferiority margin, and I'll go into much more detail of that point on the next slide.

The second very important point is to know that the study had assay sensitivity, that is, if there was a difference between

treatment and control, the study could have demonstrated that, and a study should be conducted as closely as possible to the study used to define your non-inferiority margin.

This allows one to be confident that in a particular situation, the control drug has efficacy and that there was ability of the study to differentiate between treatments.

But assay sensitivity is regarding the entire conduct of the study, from study design, definition of diagnosis, definition of endpoints and patient population, but also more elusive information, such as, was a study blind maintained, was there good follow up of patients and minimal missing data, were diagnoses made accurately and was the correct randomized therapy given?

The study should be as cleanly conducted as possible and this is often a very difficult thing to measure, especially for FDA reviewers.

The reason this point is especially important in non-inferiority trials compared to superiority trials is that a messy trial will show two treatment arms to be more similar than they actually are. This will lead to more difficult time showing superiority in a superiority trial, but will lead to more easy time, showing non-inferiority in a non-inferiority trial. Where superiority trials have built-in quality control, non-inferiority trials do not.

So, to determine a valid non-inferiority margin, we need to know how much more effective the control is, relative to placebo, and I'll refer to that as a treatment effect.

My little plot here shows the difference in cure rates. So, as opposed to Dr. Fleming's talk, where he looked at failure rates, I'm kind of flipping it and looking at cure rates.

So, you can think of the blue here

as representing plausible values for the difference between a placebo and control drug, where the diamond, say, would be the point estimate and the line would be the 95 percent confidence interval.

Margins higher than this blue line can be justified. So, in this case, a negative 15 percent margin could be justified, based on data. However, smaller margins also would be justified and may be valid, considering clinical judgment.

The green here represents plausible values for the difference between the test drug and the control and in this situation, this study would have shown non-inferiority if the margin was at say, negative 10 percent. There's no overlap between the green and the blue.

This is where problems arise, where we actually have an overlap between the plausible values of the difference between the control drug and placebo and the difference

between the test drug and control. So, in this situation, we'd need to define a smaller non-inferiority margin.

Having information from placebo controlled trials of the active control is certainly ideal to justify non-inferiority margins. Of course, we don't have that for CAP, but other historical information is available that could also be used to help justify the margins. So, those are some general non-inferiority issues to keep in mind as I go over these studies.

As I stated earlier, I'll review recent, oral-only studies for CAP. I looked at only comparative studies that were conducted within the last eight years.

There were a total of seven studies. They ranged from approximately 300 to 500 randomized subjects. The control varied, but included clarithromycin, amoxicillin-clavulanate, levofloxacin and they all closely followed the 1998 guidance for

CAP.

All seven studies were randomized, double-blind trials, designed to show similar effectiveness to the approved product, so they were all non-inferiority studies. In general, the diagnosis of CAP was based on the presence of a new infiltrate on the chest x-ray, and at least two of the following signs and symptoms, cough, sputum production, auscultatory findings, dyspnea or tachypnea, fever, elevated white blood cell, hypoxemia and note the inclusion/exclusion did vary from study to study.

Some of the studies limited enrollment to patients of fine class, less than or equal to two or less than or equal to three, and microbiologic evaluation was performed on each patient, though isolation of a pathogen was not required for overall evaluability.

Patients were assessed for outcome at the test of cure visit, which for most

studies, occurred seven to 21 days after completion of therapy. However, patients were typically seen in most of these studies, prior to that time point as well, and an earlier failure would be carried forward to this test of cure visit.

Clinical outcome was defined as a primary end point, where clinical cure is defined as complete resolution or improvement of all signs and symptoms of pneumonia and improvement or lack of progression of all abnormalities on chest radio-graphs, such that no additional antibacterial therapy is required.

The draft guidance clearly defined cure -- clearly defined failure and there was a room for improvement of signs and symptoms and in all these studies, they were included as a clinical care.

Micro-response was also defined, with eradication being absence of the original pathogen from the test of cure culture,

presumed eradication, clinical cure without a specimen for culture, persistence, presence of the original pathogen in the test of cure culture and presumed persistence, clinical failure without culture of a specimen.

The following four analysis populations were often defined in the protocol or discussed in the FDA reviews. They were the intent to treat, which included all randomized subjects. The per protocol, also called the clinically evaluable, which included all ITT subjects without any major protocol violations, MITT, which was called the Modified or Microbiological ITT, which included all ITT subjects with a treat -- pre-treatment pathogen isolated and micro-evaluable, which was the MITT subjects without any major protocol violations.

Since not all subjects were required or expected in these studies, to have pre-treatment pathogen isolated, analyses using these last two populations were usually

considered sensitivity analyses only.

Regarding which population should be considered primary, many believe that the per protocol population is the most relevant for non-inferiority studies because it removes subjects who would otherwise cloud the ability to see a treatment effect, if one actually existed.

For example, if some subjects in each arm didn't receive a minimally effective dose of therapy, including these subjects into the analysis may have the effect of making the two treatment arms look more similar than they actually are, thereby, making the ITT population a less conservative population, compared to the per protocol.

However, many others, including myself, are uncomfortable with the per protocol population alone as primary, because this population excludes subjects after randomization for reasons that may be drug related, and therefore, you may end up with a

biased population.

A population where the two treatment arms may not be similar at base line, potentially losing much of the benefit of randomization.

So, for non-inferiority trials, there are drawbacks with both populations, which is why we often consider both of them equally important in the analysis.

So, the primary analysis of these studies to assess non-inferiority was to construct a two-sided 95 percent confidence interval for the difference in cure rates, test drug minus control, for both the ITT and the per protocol populations. To conclude non-inferiority, the lower bound of both confidence intervals would need to be larger than negative 10 or negative 15 percent.

So, where did the 10 or 15 percent come from? For these studies, the margin was typically agreed upon by FDA and its acceptability was determined mainly from

clinical judgment. In general, this is how margins were selected for CAP prior to 2006.

From 2006 on, all sponsors for all indications were asked to provide data driven justifications for their non-inferiority margins.

This graph shows the percentage of ITT subjects who were excluded from the per protocol data set, for the seven studies reviewed. Note, the varying percentages range from under 10 percent to about 20 percent. These differences could be due to differences in how strictly investigators followed protocols, differences in patient populations across studies or the use of stricter criteria for entry into a per protocol population.

Reasons for exclusion varied slightly from study to study, but the following are some of the reasons used, insufficient signs and symptoms or x-ray at base line, withdrawal or loss of subjects, adverse events leading to discontinuation,

inadequate dosing, test of cure visit outside of the predefined window, indeterminate clinical outcome, use of concomitant antimicrobials not for failure and deaths not due to CAP.

It's a different version than I think you all have in your hand-out. This is an earlier version. So, these are the regions, where recent oral CAP studies were conducted. In five -- seven -- five of the seven studies included some subjects from the United States, however, U.S. subjects make up over 50 percent of the population in only two studies.

Many of the subjects were enrolled from Europe, which in this graph, includes East and West Europe and Russia, and it's in green. South America in turquoise and Canada in blue, also enrolled many subjects.

The number of countries per study ranged from three to 14 countries and the number of sites ranged from 40 to 80 sites per

study.

These are the ages of subjects. Ages of subjects ranged from 18 to 98 years old. The mean age was 46 and the median was 45, and the green here represents this middle 50 percent of the population and it ranged from 35 to 55 year olds.

This graph shows the fine scores for subjects enrolled into these studies. The y-axis gives a percent of subjects. Remember, these subjects were all oral-only -- these studies were all oral-only CAP and many were limited to certain scores of less than or equal to two or less than or equal to three. As a result and as would be expected, most subjects fall into one and two.

The percent of subjects in three or higher varied from study to study, but range from approximately five to 10 percent and note that for some of these studies, the scores needed to be calculated, based on the available data. So, they may be incomplete or

reported slightly lower than the actual score.

This slide is to give a sense of the signs and symptoms seen at base line. As mentioned earlier, some amount of signs and symptoms were necessary for inclusion into this study. I've highlighted in turquoise an outlier, when one existed.

Almost all patients and all studies had cough or sputum production. Note that many of the studies required cough and sputum production for entry. A smaller percentage of studies had fever. Ninety-eight percent is an outlier and a requirement for that particular study.

The percentage of subjects with chills ranged from less than two percent to 69 percent, but again, the 69 percent is an outlier and most were about six percent or lower.

Shortness of breath ranged from 18 to 100 percent or anything in between for these studies. Chest pain, 41 percent to 76

percent of subjects, multilobe involvement was there for 16 to 25 percent of subjects. Note that all subjects had to have a certain amount of x-ray findings, and bacteremia was rare, ranging from zero to eight percent, with Strep pneumo bacteremia being at zero to two percent of subjects.

This slide shows the percent of all randomized subjects with a pathogen at baseline. All patients should have been screened for a pathogen at entry, but as stated earlier, isolation of a pathogen was not required for overall evaluability. The percent of subjects with a pathogen varied from approximately 45 percent to 75 percent. Note that this is the population that makes up the population of subjects included in the MITT population.

In pink is a proportion of subjects with Streptococcus pneumoniae, anywhere from approximately six to 20 percent of subjects had Strep pneumo isolated.

This slide gives more details about the types of pathogens that were seen at baseline. The x-axis represents the actual number of patients with a particular pathogen, rather than the percentage, since subjects may have had more than one pathogen.

The thing to notice is just a great variety across the studies in the numbers and types of pathogens seen. Note the pink bar contains Streptococcus pneumoniae. Other common organisms are Mycoplasma in green, Chlamydia in yellow and H-flu in orange.

So, that was all the baseline information to give you in general, who was enrolled in the studies. The next slides will go over the results.

This figure reports a clinical response at the test of cure visit for the ITT population. The ITT population considers all missing data as failures. That's a primary analysis. Those sensitivity analyses were conducted using different ways of imputation.

As you can see, all showed very high success rates, all greater than 80 percent.

Here's the results for the per protocol population. Again, all very high, very similar, all greater than 90 percent.

Here is the comparative results for the primary analysis, clinical response, in both the intent to treat in per protocol. Each study is shown here twice. The green is the point estimate for the difference for the intent to treat. The pink is for the per protocol and the bars are the 95 percent confidence intervals.

You'll notice that the ITT in the per protocol analyses track very closely and notice that there's no clear pattern with which the ITT or the per protocol would lead to a larger point estimate or a higher lower bound. So, it's not clear ahead of time, which would be a more conservative analysis.

All of these studies would have shown non-inferiority with a 15 percent margin

and all but two at the 10 percent margin.

Micro-biological response is often considered an important end point and here are the results for the micro-response in the MITT population. The largest problem that we found with the micro-response is that it so closely follows the clinical response, due to the lack of ability to culture patients at the test of cure visit. It doesn't add a whole lot of information.

In all of these studies, the vast majority of micro-biological responses were presumed eradication or in the case of failures, presumed persisted and all based on the clinical response.

Finally, let's look at the rate of death. As would be expected from mild to moderate CAP, the rates are very low and are only about zero to two subjects per treatment arm. Death is an outcome used in most of historical information on CAP. It is, however, due to advances in medical care, not

really a plausible end point to use with the present studies, due to the additional measures taken when someone is failing therapy, making the rate of death in current studies non-comparable really, to the rate of deaths seen in the historical studies.

The next two slides will give just a brief summary of the IV to oral CAP studies that we've seen at the FDA. These studies were similarly designed as the oral studies. Some, however, were not blinded.

A requirement for the IV studies were that patients be newly hospitalized within 24 hours prior to enrollment. End points and definition of analysis populations and the primary analyses were all the same as for oral, and the size of the studies ranged from about 300 to 700 subjects.

Briefly, the results were that subjects were older than with the oral studies. The mean age was 56 years, with the middle 50 percent of the population ranging

from approximately 40 to 70 years. The scores were higher, signifying a more severe disease. Twenty percent of subjects had scores of three, 20 percent had scores of four, less than five percent had scores of five and the remaining 55 percent had scores of one and two.

The percent of subjects with a baseline pathogen isolated was 30 to 55 percent, slightly lower than the percentage seen with oral, and the types of pathogens really varied greatly from study to study. Approximately 20 percent had Streptococcus pneumoniae.

Eight to 10 percent had bacteremia at baseline, with four to nine percent Strep pneumo bacteremia. These rates are higher with the oral studies -- higher than with the oral studies, and clinical response rates were high, approximately 80 percent for ITT, 90 percent for per protocol. The rates of death were low, approximately two to four percent.

So, all of these studies that I reviewed here were accepted at the time as valid, non-inferiority trials, and by accepted, I mean, either the protocol was reviewed and a general acceptance of the non-inferiority margin was given or the study led to an approval of an indication for CAP. So, what's the problem?

Well, the problem is, the Code of Federal Regulations states that similarity of the test drug and active control can mean that either that both drugs were effective or that neither was effective and that the analysis of the study really should explain why the drugs be considered effective in the study, for example, by reference to results in previous placebo controlled studies of the active control drug.

Last year, the Office of Anti-microbial Products tried to gather data for a justification for a margin for CAP. We reviewed historical data, past studies,

information on the use of adequate versus inadequate therapy and Dr. Mary Singer will discuss tomorrow what we reviewed in order to try to justify a margin.

We look forward to your discussion regarding what we can learn from this information, especially from mild to moderate CAP, which is particularly challenging.

In summary, all studies that we have seen use non-inferiority trial design. The scores were mainly one and two for oral and one through four for IV. The rates of subjects with pathogens were 45 to 75 percent for oral and 30 to 55 percent for IV, and Strep pneumo ranged from 10 to 20 percent.

Low proportion of patients had bacteremia. There was a high clinical response rate and low mortality rates. However, currently, it remains uncertain if sufficient data exists to justify non-inferiority margin, especially for mild to moderate CAP, and again, we look forward to

the group's discussion today and tomorrow.

DR. COX: Thank you, Karen. Thanks for a very nice summary. We'll hold questions until after Dr. File's talk. Thank you.

DR. GILBERT: We're pleased to have Dr. File as our next speaker. Tom has a rich experience in performing clinical trials and being involved with clinical trials for community acquired pneumonia.

He is currently head of Infectious Disease and Professor of Internal Medicine at Northeastern Ohio University College of Medicine.

DR. FILE: Thank you, Dave. It's certainly a pleasure to be here and I welcome the opportunity to participate in this workshop.

I've been finding it extremely interesting and as you said, David, I've had a lot of experience as investigator in a lot of these clinical trials, not so much as a statistician evaluating the design, however,

and so, I have found -- and here's my disclosures, as you can see, but I've found my task, which is to answer these two questions, therefore, very challenging because I do want to also disclose that I am not an expert in statistics or mathematics.

As a matter of fact, when you start talking about non-inferiority margin or Delta, to me, one of the more important aspects about the Delta is that it's Greek and that's important to me, because my wife is Greek. So, she finds that very important and I must admit, it is somewhat Greek to me.

But at any rate, this is what I'm going to try to address and at least, give my personal perspective on these issues, which has already been, obviously to a certain extent, discussed and Dr. Higgins actually, in her final slide, talked about that there's question at all if there's data to be able to justify a non-inferiority margin. So, I'm not even sure how I'm going to be able to address

that.

But at any rate, let me just make some comments. They are generalized comments and then we'll go directly to these questions.

As we've already mentioned, the majority of community acquired pneumonia patients are treated as outpatients. They do have mild pneumonia and it's a very common infection and indeed, most care givers, just primary care physicians, consider themselves expert -- or at least, they know how to treat their patients with mild pneumonia.

But having said that, as we said in our initial two guidelines for community acquired pneumonia, despite extensive studies, there are very few conditions in medicine that are so controversial, in terms of management and Dr. Cox, I think you said in your initial preliminary remarks that the use of anti-microbials actually proceeded this concept of randomized clinical trials, and as Dr. Read also said in a prior, sort of evidence-based

review of community acquired pneumonia, that the hypothesis that anti-microbials are a necessary component for the management of CAP, has therefore, never been rigorously tested and I think we've already established that, because of the lack of placebo controlled trials and this is particularly the case in mild pneumonia.

However, he does go on and say that at least observations do suggest that there is some benefit to anti-microbial therapy in patients who have pneumonia, and indeed, I found this published last year by the British Medical Journal Evidence Based Statement in their handbook, where in their conclusions, they felt that antibiotics in outpatient settings, compared to no antibiotics, were beneficial.

Now, they acknowledge that this is based on consensus, but they quote that we found no randomized clinical trials comparing antibiotics with placebo or no treatment, and

such trials are likely to be considered unethical. I know we discussed this and there's some controversy here.

Then they conclude that there is consensus that antibiotics are beneficial for patients with community acquired pneumonia.

Now, we discussed some of these controversies, as far as the utility of diagnostic test. How do we actually differentiate true mild walking pneumonia from an infection which is 10 times more common, which is acute viral bronchitis, and that, I think, is a major issue when we're talking about true patients who have pneumonia, because that clinical scenario that Dave brought this morning, in a patient who has pneumonia, a positive x-ray, positive fever, other conditions there, I really think -- I mean, it would be really interesting to poll the clinicians here, but I think it would be very unlikely that we would not want to treat that patient, even though that patient has a

Fine Class 1 curb 65 score zero.

Another question that has been brought up is, what about the utility to treat a typical pathogen? So, I want to address that a little bit. There is a concern of over-use of antibiotics, such as fluoroquinolones, particularly and particularly now, in a group of patients who may be at risk for tuberculosis, and then there's this concept of mild versus moderate to severe. I think George brought that up in a discussion earlier this morning.

Actually, I view pneumonia, if it's truly pneumonia, whether it's mild, moderate or severe, as a spectrum of the same infection. It's almost like when I talk to patients who have come to me and say, "Dr. File, do I have HIV or do I have AIDS?" I say, "It's the same infection. It's just a spectrum of the same infection," and I think mild pneumonia can become moderate and can become severe.

So, the considerations that we need to evaluate are what is the benefit of antibiotic therapy, and one way to look at it is versus placebo. We've already established, we have very limited data there. There is some studies that I did find that I want to review with you, however.

So, another way to look at it is, what about effective therapy versus ineffective therapy or inactive therapy? But then that looks at this concept of resistance, and as we've already heard this morning, the clinical relevance of resistance is not well established.

I think there is some data. I say, strong and that maybe be too strong of a statement. I think there is some data which is mostly observational and retrospective that macrolide resistance can be associated with failure.

Evidence is lacking for beta-lactams, at least if you use effective beta-

lactams that it's associated with failure, and there's minimal evidence at all that there's fluoroquinolone resistance associated with failure, at least if you use the appropriate doses and the appropriate fluoroquinolone.

Then we have to consider what's the consequence of failure? I mean, we've already heard, this is a mild infection. Nobody dies of this -- or maybe less than one percent. So, we have to look at other end points. But I think there are other clinically relevant end points, as I'm going to try to show during this presentation.

As far as the end points, it would be nice if we had very objective end points, rather than just clinical impressions by the primary investigator. I'll harken back to what many have said, that for example, with HIV, we can look at for example, log-drop at 24 weeks in the viral load or what is the increase in the CD4 count, because these do have correlation with good clinical outcomes,

with these surrogate markers.

Then there's the point that we brought up earlier also as far as, what are we doing with antibiotics? Well, obviously, we want to eradicate a pathogen or effect a pathogen. But we also know, you can eradicate the pathogen, but the patient still dies.

So, in that case, we haven't shown necessarily that the antibiotic is no good, but the patient still dies because of other effects, and then there's the immuno-modulatory effect of antibiotics that was also brought up, that may confound the ability to assess the patient.

Tim already brought up the concept of what are the most likely pathogens and the one that we're most concerned about and Mike Fine did a nice study over 10 years ago, showing that the greatest morbidity and the greatest mortality is with the pneumococcal pneumonia and if you look at ambulatory patients, pneumococcus is number one and that

is what Tim showed, and I will acknowledge, if you look at some studies that look at Fine Class 1 patients, Mycoplasma may be number one, but that may be a reflection of the methodology of the study, as Tim already mentioned.

One way to look at the consequences of patients who have mild pneumonia is the failures and I'm just showing you two relatively recent studies here. The first is from Paul Iannini and Jerry Schentag's group, that -- who did a retrospect of multi-center analysis of 122 patients who were admitted with community acquired pneumonia, because they failed outpatient therapy with a macrolide and to me, this is very compelling data because as you would expect, these patients were more likely to have resistant strains because they were on a macrolide when they came in the hospital during failure. But look, 52 percent of these patients were bacteremic and there was a mortality rate of

about six percent, which is significantly higher than you would expect patients who would be treated in the outpatient setting.

Then the second study is a study from Don Low's group in Toronto, who established a theoretical model, based on linking resistance prevalence with outcomes, and so I have to acknowledge that this is a theoretical model, but it's based on an epidemiologic concept of risk difference and what they felt was, that if you had macrolide resistance rate of pneumococcus at 25 percent, which basically is what it is in North America, that you have an increased rate of death by using at least a macrolide of 1.2 percent, which is essentially double, as I understand it, what the base line would be.

Increased rate of bacteremia, 1.6 percent, increased rate for prolonged course, as much as up to six percent, if you look at the confidence intervals there.

But then we have the confounding

issue of failure that we've already mentioned, may not reflect the inability of the antibiotic to do what we want it to do, which is to eradicate the pathogen or at least, inhibit the pathogen.

If you look at the table four here, which is the bottom of this slide, it looks at a study from Tom Marrie's group where they looked at ambulatory patients who {quote} "failed therapy."

Now, their definition of failed therapy as an out patient was that they required admission to the hospital, but I want to point out that the most common reason for failure, when they really looked at this, was worsening of the co-morbid illness. It was not necessarily what they considered even clinical failure of the pneumonia itself.

So, what are potential designs for a superiority trial for a mild CAP? I do believe that if it's truly community acquired pneumonia, such as the patient scenario that

Dave mentioned earlier, that we should not be doing placebo controlled trials in that particular patient because that patient can go -- in fact, that patient, I'm sure, has a degree of chronic obstructive pulmonary disease, with a 40 pack year history of smoking.

But at any rate, because I think there is a potential for poor outcome and as Tim said, we do have good therapy for these patients.

The real problem is differentiating patients who truly have pneumonia from other respiratory infections, which do not warrant anti-microbial therapy.

I also believe that use of appropriate active controls predicts that superior results are going to be highly unlikely because right now, we've got the patients -- as we've already shown, at least for the per protocol population, where the results are well into the 90's, 90 percent

rate, at least for per protocol.

But there may be some other potential ways to do this. It's interesting, if you look at the North America approach versus European approach, at least for mild pneumonia, we do have different recommendations for empiric therapy. We recommend treating the atypical pathogens.

If you look at the British Thoracic Society Guidelines, they do not. They recommend using aminopenicillins to treat -- or to target the pneumococcus, and in fact, this is just review of our guidelines -- the consensus guidelines from IDSA and ATS, and the rationale here is, we do stratify patients according to relative risk factors and what we consider to be risks for resistance, but the point -- the rationale here is that we are targeting both pneumococcus and the atypical pathogens.

Now, I think the relevance of the atypical pathogens have, to a limited extent,

been discussed here and I know John Barlett is going to discuss them tomorrow. These studies have already -- these two meta-analyses have already been shown a couple of times at this workshop. But I think there's a lot of limitations to these two meta-analyses.

First of all, they both look at the same study, so it's not surprising that they're going to have the same results. One looks at 24 studies. One looks at 20 studies and they both come to the conclusion that there's no advantage for the -- treating the atypical -- or using an atypical regimen and no difference in mortality.

Well, there's certainly not going to be a difference in mortality, because these patients have mild pneumonia and we've already established that mild pneumonia and mortality is not going to be a sensitive indicator.

But the point I want to make as well is that most of these look at test to cure outcomes, which were like, seven to 10

days after the study drug has been completed, and that's sort of what Dr. Higgins said in the evaluation of her studies.

So, you're talking about -- and if they're going to give a seven to 10 day regimen, you're talking about three weeks after the patient presents to you, and a lot of these, as we know, infections are going to be self-limited by that time. It's conceivable that there could be a difference in more rapid resolution of the illness that we were unable to detect because of the methodology of these particular studies and indeed, the authors of the second study suggested -- in fact, this is the last two sentences of the paper, "Studies designed specifically to evaluate the necessity of atypical coverage are needed. The optimal design would be randomized, controlled trial comparing the same beta-lactam in both arms with and without an agent against atypical pathogens."

If you go back in the past, there actually are a couple of studies that have looked at the utility of effective therapy for some of these atypical -- well, I shouldn't say some, only one. That's Mycoplasma, and this is actually a well-designed trial, double-blind, placebo controlled trial done 50 years ago, mostly in South Carolina and that was a very homogeneous group of patients. There were all Military recruits. They were young.

So, I suspect, they were all port one. I think the average age was 18 or 19, but they actually treated 300 patients -- and this gets into the design issue, I guess, and then looked carefully, at least on the basis of available diagnostic methods and 109 of them had, as the sole pathogen that they were able to identify, Mycoplasma pneumoniae, because there was an outbreak of Mycoplasma pneumoniae at this Military base.

And so, if you excluded the other

patients, the other 200 that did not have Mycoplasma pneumoniae -- so, we're only looking at those patients who had documented Mycoplasma pneumoniae and they treated them either with tetracycline -- this is tetracycline treated versus placebo, and they actually used capsules that looked very similar, it was well controlled. They used IBM cards for their data in this analysis. They obviously didn't have laptops, and they found a significant difference in the rate of -- or the amount of the disease -- or at least the time to resolution of a lot of these clinical factors, which I think, can be clinically relevant to the patient.

Now, 10 years later, there was another study. This was also in Military personnel. So, young people, where they compared erythromycin, tetracycline and then, in the paper, it doesn't say -- it's either penicillin or no anti-microbial therapy, and unfortunately, they don't indicate how many of

each, but never the less, it was penicillin or no anti-microbial therapy. These patients did worse or they had a more prolonged illness than patients who received erythromycin or tetracycline.

So, I think there is some data, if you look at other end points, other than mortality or long-term end points of seven to 14 or how many days after a study drug, that there can be a benefit to antibiotics in these mild infections that I think, were port class 1.

And so, in a review paper that we wrote with many of the people in the room here, we said, "Well, maybe we need to do a large randomized controlled trial, to evaluate the difference in outcomes of the recommendations by North America and the European guidelines."

Now, in my initial slide -- and I think it's what's in your hand-out, I said macrolide versus amoxicillin. But I sort of

like the design that the Shefert study recommended at the end of their paper, "Well, everybody gets amoxicillin," because then you don't have to worry about pneumococcus. If you use appropriate doses of amoxicillin, you're going to cover the pneumococcus, even {quote} "drug resisted pneumococcus."

But then, half get placebo and half get macrolide. So, then you're -- the only issue there, however, is what is going to be the potential effect of the immuno-modulatory effect of the macrolide. But what you're going to have to do is, instead of monitoring response of 14 to 28 days, look at perhaps patient response outcomes, as we've already discussed, and do them on day one, day two, day three, day four, and see if there's a more rapid resolution of illness and we need to have accurate microbiologic tests.

I think we can do that. It's going to be expensive, but we need to use probably nucleic acid types of tests to evaluate this.

Now, as far as maybe studying new agents, certainly, double-blind, we need to do whatever we can to truly identify patients who have pneumonia and differentiate them from patients who have acute viral bronchitis. We need to distinguish what severity of illness we're going to be entering into the trial and it's probably the one's and the two's, where we really need this information, because we got more and more data about three's and four's, or the CURB-65, zero's and one's.

Because of purposes of time, I'm not going to go into to all of this, but I was really intrigued with what Michael Niederman said earlier, about perhaps using the procalcitonin or some other biologic marker to be able to help differentiate patients who truly were on antibiotics versus perhaps, those that were not.

I think we need better patient assessments, such as we've already talked about, as far as patient scoring systems.

We've already talked about the micro-biologic assessment. I really think the more we can define what the etiology is, then the more we can explain the outcomes or the results of studies, and we need to look -- instead of the standard end points here, we need to look at rapidity of resolution in morbidity, the patient-based outcome assessments that we've already mentioned and perhaps, as Mike Niederman said, the utility of what happens with biologic markers, such as procalcitonin.

With that, I'd like to maybe discuss a couple of studies that were actually designed as non-inferiority, but in which superior results, indeed, were defined and I want to correlate that with more mild pneumonia.

So, here is three studies and some of these have already been mentioned earlier, fluoroquinolone versus beta-lactam, plus or minus, macrolide. This is our study, where we showed this is a significant difference in

favor of levofloxacin, although this is not double-blind. This was double-blind and this was already mentioned.

Then this study, if you look at the standard test to cure, which was seven to 10 days after the study drug, there was no difference, but in that particular study, they did find that there was more rapid defervescence of fever and symptoms in the patients who received moxifloxacin and I only point that out because they did use a patient oriented system to evaluate the patients.

But this is -- our study, looking at levofloxacin versus ceftriaxone or cefuroxime plus or minus erythromycin, and the reason I'm bringing this up is that slightly over 50 percent of our patients were only treated with oral therapy, and so, that means they got cefuroxime, never did get ceftriaxone.

Now, if you look at overall, as I said, it's 96 versus 90 percent. I sort of

look at the failures here, not the success rates. But if you look at the patients who only received oral therapy -- so, less severe disease, and of course, this study was done before Michael Fine designed his PSI, so we don't have that, but as a surrogate, we can only look at the patients who had oral therapy only, that 95 percent of the patients were a success in the levofloxacin arm versus 88 percent here and that was statistically significant.

It's also interesting to note -- I was able to find out and I appreciate input from Susan Nicholson and Alan Fisher and Janet Peterson, from J&J and Ortho McNeil, who gave me this additional data, 11 of these 12 patients who failed did not receive macrolide. So, that may be important in that result as well.

Because of time, I'm not going to go over these other studies. This is just the study that I mentioned, looking at

moxifloxacin versus ceftriaxone plus or minus erythromycin, where they did show, in patient graded symptoms in a diary, that there was more rapid defervescence in one arm versus the other, I think, suggesting that this is a way to go, to evaluate patients as far as speed to recovery.

Dave already mentioned this study, so I'm not going to comment on that.

Finally, in trying to prepare for this, I had to do my own self-statistics 101 course. I wish I would have talked to Tom Fleming, because what he said this morning put it much more clearly than I was even considering this morning.

But just so we're on the same page, which is actually the first page, obviously, for me -- so, here's the second question here, but in non-inferiority clinical trial, using an active -- and compared to the concept, is to show that the effectiveness of the new drug compared to the active control was no less

than a predefined amount, and that's the margin of non-inferiority, which I like to refer to as Delta, and using 95 percent confidence intervals -- or limits, but we've already heard all of that.

So, what do we need to consider? Well, what is the risk associated with treatment failure, considering the severity of the disease? Well, quite honestly, in these port one's and port two's, if you're looking at mortality, there is no risk.

But there may be a difference in speed of resolution, patient benefit, as far as getting back to work or feeling better, resolution of fever. These can be important to the patient and clinically relevant.

What's the historical cure rate of the comparative? Well, we have that. We have all kinds of studies. There's well over 100 randomized clinical trials looking at different regimens in mild community acquired pneumonia, but none of them, except those

other two that I showed you, which were 40 or 50 years old, looked at {quote} "inactive arm" or a placebo arm.

And so, we have to look at the advantages and disadvantages of the drugs, or at least the comparatives that's already been brought out, that if we can have a drug, maybe that has less adverse events, more convenient dosing or adds options, which may decrease the selective pressure of resistance to other agents, that can be helpful. In fact, that reminds me -- and I forgot to mention it, when I showed the slide of our IDSA/ATS recommendations for empiric therapy for outpatients, I had in brackets, telithromycin because in our initial draft that was just about ready to go to CID, we had telithromycin as an option as well and we all felt, "Well, this is nice, because this will give us an additional option, perhaps, we'll reduce the {quote} `over-use' of fluoroquinolones."

Well, we had to sort of remove that

because a year ago, as you know, this agency was evaluating the risk benefit of telithromycin and so, that was sort of downgraded from the -- how we mentioned it within our guidelines.

And so, and this has already been mentioned, the International Conference on Harmonization, E10, that non-inferior design is appropriate and reliable only when the historical estimate of a drug effect size can be well supported by reference to results of previous studies.

Well, we've got all kinds of studies, but they're all with a control drug and we've already heard, well, does that mean they're both effective or they're both ineffective? I don't know, quite honestly and I'm not even showing this just to read it. I'm just showing -- and this is actually from Lionel Mandell's Canadian guidelines, where at the time, they reviewed a lot of the studies, which was very nice.

But as I said, there now are over 100 trials that I was able to find literature, in patients who received oral therapy for community acquired pneumonia in these {quote} "controlled trials."

So, we have all kinds of studies and at least in the most recent ones that I've shown, very similar to what Dr. Higgins said, is that the outcome is well over 90 percent in these trials, and so, if you're going to ask me, what should be the margin -- or the non-inferiority margin, then I would probably say well, probably around 10 percent. But that's totally from a novice, amateur statement.

I think, as Dr. Higgins already mentioned, there's the concern of what population you're going to use, whether it's per protocol population or ITT, and I think, you know, maybe what we need to do is just have -- I like this concept of maybe doing a superiority trial, but not necessarily reaching superiority, but if you reach where

the -- the lower end of the confidence interval is maybe minus two or minus three or whatever, but -- and you tried to do a superiority trial, but you achieved that, to me, that would show that I would feel comfortable using that drug, if I felt comfortable using the comparative drug.

So, with that, I'll conclude my remarks and thank you very much for your attention.

DR. GILBERT: Thank you, Tom and thank you for getting us back close to on schedule. We do have time for a couple of questions. Tim?

DR. MURPHY: Tom, you mentioned distinguishing community acquired pneumonia from acute bronchitis and acute bronchitis, we all know, is caused almost entirely by virus and antibiotics are likely to have no effect, and that probably accounts for much of the antibiotic misuse.

Doesn't a chest x-ray reliably

distinguish -- wouldn't that be considered the gold standard to -- if someone has an infiltrate -- a new infiltrate on chest x-ray, that's community acquired pneumonia. Whereas, bronchitis has a negative chest x-ray.

DR. FILE: Well, no, that's true, I absolutely agree that that is a differential characteristic. However, I've done a lot of these trials. I can tell you, about 25 percent of the patients -- well, now it's different because we have packs. We can look at the old x-rays.

But five years ago, when we entered patients into trials, based on {quote} "a new infiltrate," when we looked back at the old infiltrate, this was not new. This was like an old scar or whatever.

So, there's that issue about over-calling radiographic CAP versus radiographic abnormality. You can have a shadow on an x-ray and as you know, there can be a tremendous difference in the interpretation of that

particular shadow. Is this old scarring or whatever?

We found that, as I said, anecdotally, we had to drop about 20 percent of our patients, when we back and looked at the old x-ray, because these truly were not new.

Then there's the issue -- although it's not at all been studied in mild pneumonia, it's been more evaluated in patients requiring admission to the hospital, but there are those patients who -- of the x-ray is not sensitive enough, if you do a CT scan, you might find an abnormality, and we're not going to certainly measure that.

But I think -- and I didn't have time to go into this, but certainly, in the ideal design, as I said, you've got to do whatever you can to identify patients who more than likely have bacterial pneumonia that are going to warrant antibiotics, so you can see a benefit. They should have evidence of air

space disease on x-ray, I mean, as best as you can interpret.

But I think we need other pieces of information that may be helpful in differentiating patients as well, such as may be the procalcitonin or some other marker.

DR. GILBERT: Okay, I lost track of who is next.

DR. NOEL: My name is Gary Noel and I currently work at Johnson & Johnson. My question is actually for Dr. Higgins, and I don't see her up here on the dais, so I hope she hasn't left the room. But there are plenty of other statisticians here.

The concept that seems to be critical -- one of the concepts that seems to be critical in thinking about a non-inferiority trial is assay sensitivity and you talked about that, and it's talked about really in qualitative terms, rather than quantitative terms.

My question is really focused on

your review of these CAP studies and you pointed out that one of the assessments that goes into assay sensitivity is how -- again, a very qualitative term, messy the trial was.

On slide 10, I think it was, you listed some of the things that I would, as a clinical researcher, sort of point to as being messy in the trial.

In your review of these trials, are you saying that these trials were conducted in a not-messy manner or a messy enough manner that was acceptable? How can we, who are not only designing the trials, but also executing them, feel confident at the end of the day, that we've conducted a trial that has a high enough assay sensitivity to meet these criteria for non-inferiority?

DR. HIGGINS: That's a good question. It's really hard. It's hard for us to figure out what level. Certainly, one thing we rely strongly on is DSI inspections of study sites. So, that tells us a lot about

how the study was conducted.

You know, we look at the amount of missing data. We look at the inclusion criteria and were they closely followed by all the investigators? But it's --

DR. NOEL: Is that a point at which you say, "You've crossed the line here. The study is no longer sufficiently sensitive -- has sufficient assay sensitivity?"

DR. HIGGINS: Dr. Temple, do you want to --

DR. TEMPLE: Well, it's only partly messy. Part of it is very clean. To establish assay sensitivity, you need three things.

The first of them isn't messy at all. It's called in ICH-E10, HESDE, Historical Evidence of Sensitivity of Drug Effects. That is, you've got to know that the treatment you're comparing it to beats placebo regularly, and you get a number from that and from that, other things being equal, you say,

"Well, I think the effect in my new trial might be the same as the old trial."

Now, it gets messy, because now, you have to get into the constancy assumption. You have to somehow, conclude that this perfectly good effect you saw in the past persists. That's very hard. That's highly judgmental.

But what I just heard from Dr. File is that we're not anywhere close to that. We don't have HESDE yet and that makes it clean as a whistle. There's no basis for setting a margin, if you don't know what the effect size is.

I like the terms Delta 1 and Delta 2. Delta 1 is the entire effect of the drug that you believe it has -- the control drug, that you believe it has in this study, and that's based on the past. That's the entire effect.

Your non-inferiority margin can never be more than that, because if it's more

than that, then you've lost all the effect.

But in -- people are inclined to add to make it more difficult than that, by imposing a clinical judgment, by saying, "Well, I don't want to lose all of the effect. I want to lose half of it, a third of it, only a little bit of it." That's M2, the clinical judgment, and that, you pull out of the air or wherever you find judgments.

So, it's only partly messy. The first part isn't messy. The first part is, what are the data, and what everybody keeps saying is, "There isn't any data," at least not in mild to moderate. There might be for more severe ailments, which is a different question.

DR. GILBERT: Robert, I'm sorry, Dr. Shlaes was next, and then we're going to have to take a quick 15 minute bathroom break, in order to give Roger his full time. David?

DR. SHLAES: David Shlaes from Anti-Infectives Consulting, and I'm just

reflecting, once more, on the relevance of PSI, so the severity index, two trials, and I just want to point out and make sure we're all on the same page with this, is that essentially, all those patients who fall into PSI Class 1 are treated.

So, the mortality numbers that we're seeing are all in treated patients. Probably none of those patients are not -- or at least not intentionally not treated.

So, those are all treated patients, I would presume.

DR. GILBERT: Historically, I think we agree with you. We didn't have PSI earlier, when some patients were in treatment.

All right. So, Dr. Echols is going to lead off, and we're going to start promptly in 15 minutes. Thank you all very much.

(Whereupon, the foregoing matter recessed at 2:33 p.m. and resumed at 2:43 p.m.)

DR. GILBERT: So, the whole idea of

this workshop was to get all the ideas out on the table, and that certainly includes our colleagues in industries. So, the next presentation is by Roger Echols, Chief Medical Officer at Replidyne, Clinical Trial Design for mild to moderate community acquired pneumonia, and this is a reality check.

DR. ECHOLS: Thanks, David. Thank you very much and I really want to thank all the organizers of this meeting for really both, putting the meeting together, which I think is really critically important to get some of the diversity of ideas out on the table, but also to allow an industry perspective, and I'm very pleased to try to provide that.

Some of this may be a little redundant, but I hope it's complementary and not boring to you. But just from, again, our perspective, where did we get -- how did we get to where we are?

So, we have not only the recent

guidance from October 2007 on non-inferiority margins, but we have other historical issues about non-inferiority.

But the non-inferiority guidance of October 2007 really made two points. One, that it's not possible to define a non-inferiority margin for active control in non-inferiority studies in acute bacterial sinusitis, AECB, or acute otitis media.

But the second point and the one which we're addressing at this meeting is how to determine an NI margin for other indications, to -- in the words of the guidelines and ICH, is to ensure that there's adequate scientific rationale for the effect size of the active control and the proposed non-inferiority margin. That's what we'd like to try to get to.

The FDA non-inferiority guidance refers sponsors to the ICH E-10 document for further guidance. It is important to understand that the ICH E-10 is a general

guidance for industry and not specific to anti-bacterial drugs used to treat acute infectious diseases.

It is clear from the E-10 document that for demonstration of efficacy, superiority trials, either placebo or active controlled, are preferred and that non-inferiority studies are problematic due to the difficulty in determining the non-inferiority margin.

There is a need to demonstrate the benefit of active control over no treatment, referred to as M1, before one can determine the actual NI margin or M2.

This process should be based on statistical reasoning and clinical judgment, although it's not clear where statistical reasoning ends and clinical judgment is allowed to contribute.

The E-10 stresses that historical basis for M1 should be determined from clinical trials where the patient populations,

the outcomes measured, and the concomitant therapies should be similar to the proposed NI studies.

Finally, there is a concern about approving new drugs which may be less effective than the control drug, even if only by a small margin and the possibility of bio-creep, if this process is iterative.

Not withstanding the statistical reasoning that places such inherent value on superiority trials, I think it's important to share with you some real world experience regarding placebo controlled superiority trials in indications such as AECB and ABS.

Bayer has been conducting a placebo controlled trial in acute bacterial sinusitis in North America, which is now in its fourth winter respiratory season. Their goal is to get 117 micro-biologically evaluable patients.

Our own placebo controlled trial in AECB has been enrolling subjects for more than

two years. As difficult as patient enrollment has been at the site level, we've been sobered by the resistance to placebo controlled trials by international ethics committees and ministries of health. These organizations, which function under the same ICH guidelines as the FDA, have a far different view on the need for superiority trials.

The most common reason for rejection is the fact that the placebo controlled trials contradict established treatment guidelines for the indication being studied.

In addition, some European countries, while accepting the rationale of establishing definitive efficacy versus placebo, never the less, find a study without an active control of no value and hence, unethical.

This recent experience in infections far less severe or serious than CAP provide ample evidence against placebo

controlled trials for even mild to moderate CAP.

What is mild to moderate CAP? While this two day workshop is divided into the discussion of mild to moderate CAP versus more severe CAP, there really is no good way to separate these two indications. There is little scientific evidence to suggest the microbial etiology is significantly different.

From a regulatory perspective, oral therapies are usually excluded from labeling for severe infections, although pharmacodynamic parameters would not support this distinction for drugs that are highly bioavailable.

One only needs to look at clinical practice in other countries, to realize that the use of parenteral versus oral therapy in non-ICU patients has more do with hospital reimbursement than medical science, and while scoring systems are predictive of overall mortality, they have more to do with age and

comorbidities than the actual severity of the acute episode of CAP.

Several years ago, I was directly involved in a large clinical program for an antibiotic which ultimately was not approved for marketing. This program included seven CAP trials conducted globally, which enrolled over 2,200 patients.

All trials characterize patients based on Fine Score, or PSI score. I still refer to it as Fine Class.

Two trials included only Fine Class 1 and 2, treated orally, with orally administered drugs on an ambulatory basis. Two trials involved only hospitalized subjects initially treated with intravenous therapy, and the other trials were flexible with regard to location and route of administration.

Sixty-three percent of subjects had either a typical or an atypical pathogen identified. However, nearly 25 percent had mixed infections, usually a typical and an

atypical pathogen.

In 2003, I analyzed these pooled data CAP studies -- from the CAP studies to determine whether there was any difference in the pathogens based on Fine Class. These data were presented at the annual IDSA meeting in San Diego in 2003.

What we found is that there was very little difference in the specific microbial etiology across Fine Classes. Strep pneumoniae was the most common typical pathogen for all groups, followed by Haemophilus influenzae.

Among the atypicals, only Mycoplasma pneumoniae appeared more frequently in Fine Class 1, relative to the other Fine Classes.

We concluded that the etiology of bacterial pathogens was not different across Fine Classes, and therefore the specific microbial cause of CAP is not the reason for differences in mortality observed in the Fine

Scores.

Recently, our group at Replidyne conducted a detailed review of various summary basis of approvals, available on the FDA website or though Freedom of Information. We selected CAP trials since the early 1990's where the disease was not severe and where a systematic search for both typical and atypical pathogens was prospectively conducted.

In a pool of 5,025 evaluable subjects, 55 percent had no microbial etiology identified. In the 45 percent who had an identified pathogen, about two-thirds were typical bacteria. Strep pneumoniae was the most common typical pathogen, followed by Haemophilus influenzae.

We also conducted a literature review over the past decade of epidemiology studies which included more than 7,400 well characterized subjects with mild to moderate CAP. Again, Strep pneumoniae was the most

common typical pathogen and Mycoplasma pneumonia was the most common atypical pathogen.

Now, while the methodology of the patient definitions may differ between these various sources of information, the similarity of the results strongly support that -- the frequency and importance of Strep pneumoniae and other typical pathogens in mild to moderate CAP. Thus, is it appropriate to consider CAP as a continuum of disease of varying severity and not as a separate disease from severe CAP.

In order to conduct a scientifically rigorous non-inferiority trial in CAP, we need to establish the benefit of antimicrobial treatment versus no treatment. While this cannot be achieved through contemporary placebo controlled clinical trials, it is clear that specific antimicrobial chemotherapy, first demonstrated by the sulfonamides, had a profound impact on

patient mortality due to Strep pneumoniae.

Evans and Gainsford reported on a reduction of mortality from 27 percent to eight percent in two cohorts of subjects with lobar pneumonia. Although the study was not randomized in a manner we would find acceptable today, it did have a contemporaneous and well-matched control group.

Following the sulfapyridine dosing recommendations of Evans, Flippin, et al. reported on a cohort of 100 cases of documented pneumococcal pneumonia admitted to several Philadelphia hospitals.

In addition to the four percent mortality rate, they reported in detail the dramatic clinical response observed in their patients. Fully 83 percent had a substantial drop in temperature in the first 48 hours.

While sulfapyridine chemotherapy and penicillin clearly had an impact on mortality, using mortality as an endpoint in

CAP clinical trials for a new drug is not appropriate or feasible.

Can we ascertain the benefit of antimicrobial therapy on clinical response based on published historical data? Well, while Flippin described clinical response in a cohort of sulfapyridine treated subjects, there was no control group.

In examining the pre-serum and pre-antibiotic data, we discovered an amazing text by Bullowa which details the natural course of clinical resolution in 662 patients with serotyped pneumococcal pneumonia.

This cohort of survivors received neither the serum therapy nor anti-microbial therapy, and from this large data set, it is clear that spontaneous resolution does not occur rapidly.

Crisis, the term used to describe the dramatic drop in fever and clinical improvement, does not occur before 72 hours. It usually takes seven to nine days, and in

fully 10 percent of his patients, initial clinical resolution in survivors did not begin before two weeks.

Bullowa's observations were supported by Osler in his 1910 version of Principles and Practice of Medicine, and in contrast with the 1942 edition, written by Christian, where it is expected that a rapid clinical response would occur within 24 to 48 hours following treatment with sulfapyridine.

What about clinical response in present day circumstances? Again, we looked at the many CAP trials conducted since the mid-1990s and focused on those subjects who were clinically or microbiologically evaluable. The data shown here includes more than 3,600 clinically evaluable and 1,180 microbiologically evaluable subjects. Again, these data are FDA reviewed data.

In the subjects with mild to moderate CAP, the clinical response of cure or improvement was nearly to 92 percent, similar

to what Dr. Higgins presented, and slightly higher for the microbiologically valuable population. It made no difference whether the subject's pathogen was a typical or an atypical organism. What is striking from these data is the consistency among trials.

While the clinical response results from the summary basis of approvals represent a dichotomus variable at a specific point in time post-treatment, others have looked at time to response as a continuous variable.

I've described the Bullowa data of spontaneous resolving cases of pneumococcal pneumonia. Petersdorf, in a study previously described by Dr. Gilbert, conducted a randomized controlled trial of penicillin plus aspirin or placebo to determine the added benefit of antipyretic therapy in pneumococcal pneumonia.

He designed a scoring system of clinical signs and symptoms which were monitored on a daily basis, and while he found

no benefit of aspirin beyond the first 24 hours, he did document the rapid improvement in signs and symptoms in patients treated with penicillin.

More recently, in separate studies, Halm and Menendez characterized the time to clinical stability in hospitalized patients with CAP, and while the median time of three to four days is relatively short, both of these trials lacked microbial diagnoses, in essence relying on clinical diagnosis for patient inclusion.

Finally, studies by Dean and Torres, also discussed earlier, have prospectively monitored time to response as part of comparative non-inferiority trials in mild to moderate CAP. While both trials used respiratory quinolone versus a macrolide or amoxicillin, despite the use of a validated patient oriented questionnaire, developed by Lamping and presented by Dr. Gilbert, these instruments were totally unable to distinguish

between two very different active treatments, as illustrated in the following slide. This is a representation of table five from the manuscript and was presented earlier by Dr. Gilbert.

The CAP 2000 study compared moxifloxacin versus standard of care, which included either clarithromycin or amoxicillin or a combination of both in a double-blinded trial in ambulatory CAP patients.

While knowing the time to response may be of interest to sponsors and clinicians, such an analysis is not suitable for regulatory approval in CAP, since there is no evidence it can distinguish superiority between active therapies and it would be even more difficult to justify a non-inferiority margin and establish a study sample size based on a time to response outcome.

What can we conclude from these clinical trials and other historical clinical data sets? First, that clinical response in

bacterial pneumonia treated with an appropriate antimicrobial drug is rapid, certainly when compared to spontaneous resolution in those subjects fortunate to survive pneumococcal pneumonia.

Subjects enrolled in clinical trials who have not improved clinically in 72 hours are usually considered treatment failures and re-evaluated for alternative diagnoses, complications such as empyema, and the need for alternative antimicrobial treatment.

Second, there is little evidence to suggest that a time to response outcome variable would be better able to discriminate between two active treatments in CAP. This is not surprising, since we know that the clinical response has more to do with host factors and disease severity than the specific drug-bug interactions.

What are the prospects of achieving superiority in an active controlled trial in

mild to moderate CAP? The preponderance of data would suggest that this is unlikely to occur, even when stacking the deck, as in this study by Petitpretz, et. al., where the respiratory quinolone was compared to amoxicillin in a study designed to enroll subjects with penicillin non-susceptible Streptococcus pneumoniae.

Even with the added activity against atypical pathogens, looking at the sub-set -- and looking at the sub-set of PRSP, amoxicillin was not inferior to moxifloxacin. There are at least five additional trials comparing a respiratory quinolone or a macrolide against amoxicillin, all of which failed to demonstrate superiority, which would be expected on the basis of in-vitro-susceptibility.

There is, however, a trial showing superiority of levofloxacin, when compared to a regimen of ceftriaxone followed by cefuroxime or cefuroxime alone. This was

discussed in part by Dr. File. Actually, the publication, Tom is the lead author.

The subjects in this trial were largely defined as having mild to moderate CAP. More than half were treated entirely as out-patients, and this meant that half of the cephalosporin group received only cefuroxime.

The data here that I'm presenting is the medical reviewer of the FDA's data, not the data from the publication.

Based on the FDA medical reviewer's assessment, levofloxacin was superior to the cephalosporin regimen, for both the clinically evaluable, where the difference was 12 percent, confidence intervals here, as well as the microbiologically population, where the difference was 16 percent, and you see the confidence intervals.

It is important to note that cefuroxime is not approved for CAP in the United States, and the dose used, 500 milligrams twice a day, is one-third the dose

recommended in Europe for initial treatment of CAP.

Thus, while cefuroxime, as utilized in this study, may be considered sub-therapeutic, it is still likely better than placebo. This study is important because it demonstrates the clinical and microbiologic superiority of levofloxacin in a contemporary clinical trial, a study which was carefully reviewed by the FDA and which allowed a superiority claim in the package label for levofloxacin.

The observed differences of 12 percent for the clinically evaluable population and 16 percent for the microbiologically evaluable population under estimates the real benefit of M1 of levofloxacin versus no treatment, given the likelihood that the cephalosporin regimen had some treatment effect. The study is contemporary and provides substantial microbiologic documentation.

Although this study has not been reproduced, we do believe it provides one approach to justifying a non-inferiority margin in mild to moderate CAP. Specifically, it supports an NI margin of 10 percent for a clinically evaluable population and 15 percent for the microbiologically evaluable population.

Another approach at justifying the NI margin is a bit more convoluted but takes into account the historical Bullowa data for spontaneous clinical response in patients receiving no therapy for documented pneumococcal pneumonia. These data represent the best placebo group, where clinical response and not mortality was the outcome measured.

If we accept the premise that spontaneous clinical response does not occur within 72 hours, whereas a lack of clinical response in that same time frame would be considered treatment failure in the antibiotic

era, then the benefit of antimicrobial treatment is quite large.

To define M1, we can use the observed clinical response for microbiologically evaluable subjects derived from recently approved drugs of 93.8 percent for mild to moderate CAP, and then take the lower boundary of that 95 percent confidence interval, which is 91.3.

We then multiply this times the proportion of enrolled subjects expected to have typical bacterial pathogens, here estimated at 35 percent. In other words, we're excluding the possibility of the antibiotic having any benefit in patients with either no diagnosis or atypical organisms, a very conservative estimate.

This determines the M1 of 31.9 percent. To then determine the M1 margin -- to then determine the non-inferiority margin for future CAP trials or the M2, we conservatively take 50 percent of M1 or 15.9

percent for the microbiologically evaluable population.

Given the fact that the Bullowa only include documented bacterial pneumonia, we cannot estimate the non-inferiority margin for clinically evaluable population.

Furthermore, while the strength of this estimate lies in its detailed documentation of the historical data, we must recognize that CAP is not caused only by the pneumococcus and that supportive medical care has improved greatly since the pre-antibiotic era.

I've presented data for both clinically evaluable and microbiologically evaluable subjects. The distinction is important since what population is primary in the study analysis will determine the study sample size. The FDA prefers, as Dr. Higgins pointed out, two or co-primary populations in their analysis of non-inferiority trials. In the past, these have been the clinically

evaluable or per protocol and the ITT populations.

Currently, the FDA is requesting the CE population and the MITT, defined as ITT subjects with a microbial etiology as the co-primary populations.

Since the MITT population represents a much smaller subset of subjects, estimated here to be 30 to 35 percent for typical pathogens, than the CE population, a study previously sized to show non-inferiority of 10 percent with 484 subjects enrolled would now require nearly 1,200 subjects, should the same 10 percent non-inferiority margin be applied to the MITT population.

However, if the non-inferiority margin applied to the MITT population was 15 percent, the sample size would be 556, a number much closer to the 10 percent margin for the CE population.

This brings us to our proposal for a non-inferiority trial in mild to moderate

CAP, where the co-primary populations are the clinically evaluable and the MITT. However, the NI margins for the co-primary populations are different.

Based on the levofloxacin trial and -- versus ceftriaxone and oral cefuroxime, we feel a non-inferiority margin of 10 percent for the clinically evaluable population is justified.

Furthermore, based on both this levofloxacin study and the historical data in documented bacterial pneumonia, a non-inferiority margin of 15 percent for the ME or MITT population is justified.

With co-primary analysis, a sample size for one study would now be 618, up from 556. Assuming two trials are required for approval, the total number of CAP subjects required would be 1,236.

While this would not -- while we would not suggest pooling these studies in order to add an additional hypothesis to test

for, it is of interest to see that there is adequate power, more than 85 percent, to show a non-inferiority -- to show non-inferiority using a 10 percent NI margin for the pooled MITT or ME populations.

Let me summarize what I've tried to present, from an -- as an industry perspective on clinical trials in CAP. First, we think the evidence supports the fact that CAP represents a continuum of disease, not separate entities, dependent upon some distinction for patients able to be treated with oral antimicrobials.

Second, while recognizing the statistical reasoning for superiority trials, neither placebo controlled nor active controlled superiority trials in CAP, even mild to moderate CAP, are feasible. Looking for alternative outcomes, such as time to response, is not likely to alter that fact.

Third, that non-inferiority margins for mild to moderate CAP can be justified

using clinical judgment and statistical reasoning. While there is a need to take some license with the preferred methods found in the ICH guidelines, these are, in fact, guidelines and not statutory requirements.

Fourthly, the question is not just the absolute non-inferiority margin, but what populations will be included in the primary analysis. The impact of this decision will greatly influence sample size and thus the feasibility of trials.

While CAP is not an important commercial objective, it is considered the anchor for other respiratory tract infections.

Finally, I'd like to stress the need for regulatory clarity and a definitive transparent decision on the questions of study design before us. Without regulatory clarity and an acceptable path forward, new investment in antimicrobial drugs will diminish.

I'd like to thank members of the Replidyne team, especially Glenn Tillotson and

Bob Tosiello for their contributions to this presentation. Thank you for your attention.

DR. GILBERT: Thank you so much. We are open for comments and questions. Yes, Daniel?

DR. MUSHER: Could I ask you to show the figure in which there was a -- measuring certain sign and symptoms, I don't know which they were in pneumonia and one patient -- one set of patients, it was either the quinolone or maybe the one with the amoxicillin and clavulanic acid, and you showed that the improvement was very consistent in the two groups and from that, you said that you didn't think this kind of a study was a valid study.

I guess my conclusion would be, I think it's a very valid study and these two drugs happen to do approximately equivalently.

DR. ECHOLS: I wasn't -- what I was trying to say is that the sensitivity of this barometer, this patient oriented questionnaire, is not going to distinguish

between two active treatments.

DR. MUSHER: But it might, if there weren't -- if they weren't such good treatments, it might, no?

DR. ECHOLS: Then you're --

DR. MUSHER: I'm missing the point.

DR. ECHOLS: Well, the point is, is that to try to do superiority trials in active control trials, even if you tried to use a different outcome measure, you're not going to succeed.

DR. MUSHER: And the reason you wouldn't is because these two drug regimens are so -- really fine, you're not going to get one superior?

DR. ECHOLS: If you really use -- if you use a truly inferior drug regimen, the studies --

DR. MUSHER: It'll pick that up?

DR. ECHOLS: -- will not be conducted. They'll be considered placebo. They won't fit with guidelines.

DR. MUSHER: Okay, I guess my way of looking at it is, this is the kind of study that we should be doing, because I think it gives us pretty good insight into how good the drugs are.

DR. ECHOLS: I think the information --

DR. MUSHER: The 21 day outcome isn't so helpful.

DR. ECHOLS: The information gained from a PRO or a patient questionnaire is valuable. I'm not disputing that. But it's not going to be a tool that will allow a study that was otherwise a non-inferiority study to also become -- to all of a sudden become a superiority study.

DR. MUSHER: Okay, I'm sorry, I certainly --

DR. ECHOLS: That's the point I was trying to make. I'm sorry.

DR. GILBERT: Yes?

DR. TEMPLE: You may have said this,

and I may have had a postprandial failure here. Your estimate of what the untreated response in, say, three days or something, was essentially zero.

DR. ECHOLS: Essentially zero for documented bacterial disease.

DR. TEMPLE: Okay. So, if that's the case, then any response is by three days or something, must be attributable to the drug and you have a rock solid control rate. But where does -- can you say again, where that view that nobody is better by three days comes from?

DR. ECHOLS: It comes from the Bullowa data. Actually, when I say nobody, it's a slight exaggeration.

DR. TEMPLE: Okay.

DR. ECHOLS: One-point-three percent, and this is after hospitalization, so we really don't know how they've been ill. But only 1.3 percent, and they took these 662 cases and documented their clinical course on

a day-by-day basis.

DR. TEMPLE: Okay. So your fundamental contention is that if you look early and look for early response, you have something where the spontaneous rate is essentially zero?

DR. ECHOLS: Yes.

DR. TEMPLE: It's sort of historical data, okay.

DR. ECHOLS: So, I'm not -- and this was a cohort of patients, all of whom survived. So this is not a mortality endpoint study. This is really as best, I think, you can find as a placebo group with a documented disease.

DR. TEMPLE: I think people will conceivably raise issues about whether, in the modern world, we would do better than that. But that is an impressively low number.

DR. O'NEILL: Yes, what I don't understand is, that's a length by a sampling problem. It's conditional on only those who

have survived. So, if you sort of had a time zero and sort of looked at three days from time zero, whatever that might be, you'd have a different answer here.

So, I'm not so sure it's as impressive as you're making it out. But it's 1937. Maybe everybody ought to get a copy of the book.

DR. SPELLBERG: But just to clarify --

DR. ECHOLS: I've provided the chapter to the agency earlier this summer.

DR. SPELLBERG: But if you only look at survivors, that means the ones you're not looking at are the ones that died. So, the response rate should be worse --

DR. ECHOLS: They are.

DR. SPELLBERG: -- in that. So maybe I missed your point.

DR. ECHOLS: The point was try to construct --

DR. SPELLBERG: I wasn't -- sorry.

I understand your point.

DR. ECHOLS: I was saying to Dr. O'Neill --

DR. O'NEILL: Yes, I'm trying to reconstruct what this would have looked like if you had a cohort of a treated and a non-treated that started from time zero, whatever that time zero was, and then looked at three days from time zero to see whether there was any response. That's essentially what I was -- and in my mind, conceptually thinking about the way this was, the time zero was in survivors, those who had passed through some time and then looking at time from -- where ever -- time zero, and I'm just not sure what time zero is.

DR. GILBERT: And right, we'll get to you in a second, and Roger, isn't this data, the Bullowa data, also consistent with the Gold and Austrian data, which was no pneumococcal disease, where treatment, whether it was serum treatment or placebo or

penicillin, there was no effect on the first three days of their survivor sheet?

DR. ECHOLS: Yes, and that's part of the basis why we've often excluded people from trials, if they died in the first 24 or 48 hours, because they were going to die no matter what.

But the key thing to remember with the Petersdorf, which is based on Max Finland's data and everything, those data were in bacteremic pneumococcal disease, and clearly the mortality rate was very substantial in that group.

DR. GILBERT: Yes, Dr. Rex?

DR. REX: John Rex again from AstraZeneca. Whacking your head into an immovable object is remarkably clarifying. Before your talk, we were listening today that there are no placebo controlled CAP data. Even when we do have data, they look at outcomes after several weeks, 10 days of therapy, 10 days to test of cure.

The general background, medical care has changed. So Dr. Fleming's assumptions of constancy are completely unavailable to us. The occasional superiority in an active control might be just by chance, and yet we also believe that bacterial pneumonia is a real entity.

Then, you've -- there's a magnificent review. You have pointed at a pragmatic solution to an absolute box that we're backed in to, because the pristine science is very clear. I really enjoyed the talks this morning. It is really clear what the pretty science would look like.

But let me remind you of the question that I asked earlier today. It takes years for industry to create a new drug. We have to be able to see how to get from hither to yon. You cannot wait until the day you want the new drug to say you want it. You have to tell me about 10 years in advance. We have to have pragmatic solutions now that will

allow us to develop drugs in a setting where resistance isn't everywhere. And so, I think what you propose is brilliant.

I wanted to add, I think, three things to your set of observations that support the pragmatic things you pointed at. The first one is, I want to remind everybody that we do have some other hints about activity.

While we may not always know how to measure resistance, for at least for the macrolides, when we see strong, erm and meth-based resistance together, resistance rates go up. So, whatever you want to say about that -- I'm sorry, failure rates go up, excuse me.

When we see true macrolides, strong macrolide resistance, the clinical failure rates go up, suggesting that without the resistance, the macrolides actually are doing something, and that feeds into Dr. Higgins's summary, where all the drugs looked about the same. The macrolides are doing something some

of the time. So, I leave you to work them out.

We see sputum culture conversion, suggesting that we're doing something to microbiology. Now, John Powers has often reminded me that clearing the bug out of the sputum is not the same as curing the patient, but it's kind of in the right direction.

The other thing that we've not talked about at all is the fact that infection is blessed with the absolute best pre-clinical models of any disease area. Superb. We kill the bug in the test tube. We kill the bug in a mouse. We kill the bug in a mouse's lung. We kill the bug in a mouse's thigh. We kill the bug in George Drusano's hollow fibers. We demonstrate -- and we demonstrate with that, how much drug you've got to have for how long at the active site to do something to the bug.

So, you put all that together with your very pretty summary, and this business about crisis taking some days, that's embedded

in all the literature of the 18th century.

Fascinating observation. That was a lovely talk. I just wanted to say that again, and I wanted to add my list of a few things that, again, support the notion that something really is going on with the current drugs.

DR. GILBERT: I think Tom has a comment for you, actually.

DR. FLEMING: Actually, it's more on the entire presentation. I think the key slides, if I'm following this, are slides 17 and 18 -- and I'll just look very briefly with you at slide 17. I think you concluded that the margins that would follow from slides 17 are 10 and 15 percent, and I would say, that's a real reach to conclude that based on single trials that would show something here and with lower limits of six to seven percent and no adjustment for preserving half the effect and with all the compelling arguments that we certainly, don't want to be giving up, we

can't do placebos. We surely don't want to be losing a substantial part of the effect of the active comparator.

But the real slide is slide 18, and I would -- there's a whole lot that needs to be better understood before this really leads to a strong argument for the margins.

The basis for this, as I'm understanding, is essentially the argument that there's 35 percent of the population in which you would have essentially no success, and, yet, we're getting 93 percent success overall.

And so, essentially, what we're concluding is that there must be at least an M1 or a delta of 32 percent, and are we -- essentially, what you're trying to argue is that there is 35 percent of our population of these studies in which in the absence of these standard therapies, they all would have done essentially zero in terms of the response where we got 94 percent or 93.8 percent.

DR. ECHOLS: Or they would have spontaneously been cured, and again, these are post-treatment test of cure assessments --

DR. FLEMING: So, the 93.8 percent comes from what we were looking at in all the studies that Dr. Higgins showed, correct?

DR. ECHOLS: Dr. Higgins, in our analysis, which sort of came from this, this was our summary, we included more studies --

DR. FLEMING: Right.

DR. ECHOLS: -- because we went back to 1995 or so --

DR. FLEMING: And that's at what time period?

DR. ECHOLS: This is the typical test of cure, end of treatment, a week or so after treatment. So, it's --

DR. FLEMING: It's not three days. Now, that's only a partial issue, that it's not three days, because I think what you're claiming is, at three days, where it's 1.8 percent, but by later on, I thought you said

maybe it's only going to be up to 10 percent or so.

DR. ECHOLS: But at this point, again, Dr. Higgins iterated this, is that if someone is a failure by day three, that failure response is carried forward.

DR. FLEMING: So, what you're -- but basically, the bottom line that you're trying to argue here is that you're going to have a third of your population in the absence of these standard therapies that you're going to believe will have non-successes in the context of the non-inferiority trials that are being done today, and that's an incredibly strong assumption. It seems highly implausible as well that everything that we do changes a 60 percent response rate and makes it almost identical, 91, 92, 93, et cetera.

When I look at the time to prevent analysis that you do, showing no difference, and I'm looking at Dr. Higgins's analyses that show everything looks almost exactly the same,

I'm really skeptical that that's a scenario where everything is doing relatively little, because if everything is having a huge effect, isn't there any suspicion that everything we do has exactly the same huge effect?

This is an -- before one would take this type of analysis as a basis to justify a 15 percent margin, there's a whole lot more understanding that needs to be in hand as to the relevance of what you're assuming, i.e., that you're assuming we can say reliably that a third of the patients in all of these trials would have been failures on this clinical cure rate assessment, had we not offered them this active intervention.

DR. ECHOLS: I think the third -- the number, the third here, that I'm using, is really to be more conservative and not take any credit for clinical response in patients that had either typical -- excuse me, had atypical organisms or no organisms.

So, the third here represents only

those -- that proportion of patients from these large databases and basically, confirmed by everybody else's review today, that about a third of the patients have -- or higher, because of our inability to diagnosis, a third of them have typical bacterial disease --

DR. FLEMING: Bottom line though, what you're saying, in order for this argument to fly, is that we can reliably believe that the patients that are in the studies that Dr. Higgins summarized, all of whom had 92, 94 percent response rates in Bob Temple's per protocol analysis, would have had, to justify this margin and to get this M1, a response rate in the neighborhood of 60 percent in the absence of the use of that standard of care intervention.

DR. TEMPLE: I don't know how surprising that is, if you think you understand what bacteria are and how they respond to things that kill them.

If someone told me that for a

urinary tract infection, where I was only looking at organisms that were sensitive to the -- to the antimicrobial, that wouldn't surprise me particularly. It doesn't surprise me that things that they're sensitive to kill them at about the same rate, whenever you look at how killed they are.

DR. FLEMING: Well, presumably, we're looking at a clinical cure endpoint, not a microbiologic endpoint. We're looking at --

DR. TEMPLE: No, no, I know.

DR. FLEMING: -- clinical resolution of symptom, et cetera.

DR. TEMPLE: I understand, but I do have the belief that bacterial pneumonia has at least something to do with the growth of bacteria, and -- well, you know, I'm a Bayesian; I've got priors, and I'm embarrassed about it, you know.

DR. REX: Let me remind us all that the doses and the exposures that were chosen were chosen because it looked as if that ought

to have an effect on the bacteria.

So, it's not -- these are not arbitrarily chosen regimens. They may not have all been worked out with the same lovely degree of pharmacodynamic work that we know how to do now. But they were chosen to get some kind of an exposure in the right organ at the right dose.

DR. TEMPLE: No, I totally agree, that's why I find it plausible compared to most other settings, because if you think of the human being as a big test tube, in which you're putting antibacterials in at a value that kills in the test tube, it ought to kill them there, unless there's some inaccessibility or something weird, and that's why you need to do trials.

So, I don't find it as totally astonishing as the consistency at first appears. Also, it's not perfectly consistent. It goes between 80 and 90 and stuff.

DR. FLEMING: Well, in fact, it's --

if you look at the studies Dr. Higgins put forward, it is very strikingly consistent. There is, in fact, i.e., slide 17, is really a very uncommon scenario where you're going to see something even of the magnitude of 12 percent.

So the issue is not is it plausible that these interventions are all having a microbiological effect. The issue is, if we're saying we believe that in fact, the true clinical response would have been only 60 percent, what is implausible is that everything that you do gives the same clinical outcome of 90, to 92, to 93 percent, and in fact, the time to resolution is the same as well.

DR. TEMPLE: It would be good to go back and look -- I mean, if you found that in strep throat or in something relatively simple like a urinary tract infection, you would not -- unobstructed urinary tract infection, you wouldn't be that surprised because it's like a

test tube.

DR. SPELLBERG: But it's also like pre-clinical data, and I think Dr. Rex made that very important point. The subtleties between the ability of different antibiotics to kill organisms -- and this is true whether you're talking bacterial or fungal, are -- they are probably important to some degree, but over and over again, we've had inability to show differences, either microbiologically or clinically, whether we're using -- or static therapy.

So, I think pre-clinical data also supports the concept that if you kill bacteria, you tend to get similar clinical response rates or pre-clinical response rates.

DR. POWERS: So if it's all about killing the bug, why aren't more potent drugs superior to less potent drugs, clinically, on important clinical outcomes to patients?

DR. SPELLBERG: That's exactly -- I mean, there has to -- there's a threshold

effect. There's a degree to which if you can inhibit growth of the bug, it doesn't matter whether you kill them three logs in 24 hours or two logs in -- there may be subtle differences, but it's very difficult to show.

DR. POWER: Right, and there's this little thing called the human immune system, which is also working very hard to kill the organisms as well, and the entire question is, does giving a drug, which may help your immune system get rid of the bugs, does it make the person get better faster or decrease mortality, and that's the whole question.

DR. TEMPLE: But John, the underlying premise here was that over a three day period -- I have no basis for knowing whether that's believable or not, but that's what we're told. Over a three day period with -- in people like this, you don't see much benefit.

Now, maybe in two weeks, the immune system will kick in and they'll get better and

all of that. But over three weeks, you don't.

The other thing, it seems to me worth remembering, is there are some data that say if you have a PPLO organism, you do better if you include a drug that actually goes to that organism. Well, that's some information about one kind of pneumonia, and I guess no one will let you do the trial in which you take people with the resistant organism and randomize them to the thing they're resistant to or to something they're not resistant to.

DR. POWERS: We don't know that at the start of the trial.

DR. TEMPLE: No, no, but what about people who, at some point, have proved resistance and are still doing badly at, say, four days? Would anybody let you do the test that would be informative, which is to randomize them to something that they're not resistant to or to randomize them back to the thing they failed on?

Well, probably no one will let you

do that, but that would answer the question, too, if anybody would.

DR. POWERS: I think that there is -- we've sort of taken this number of 1.3 percent and run with it. Having reviewed all of this information as well, there are -- this sort of violates every principle of the constancy assumption, which is, I think, is what Tom was getting to.

It also doesn't address the question -- and Roger, let me get across that E-10 applies to antibiotics, just like it does to everything else. I think we've actually gotten to this point because we think somehow antimicrobials are different and none of this stuff applies. But it also -- E-10 talks about reliable and reproducible benefit.

Now, when you look across other things and other studies by Davies, and Bullowa has actually got a bacteremia study, and a bunch of other people that have looked at these things in the 1800's, it's hard to

reproduce that, actually.

So, the question that comes up is, you found one, where is the others that actually confirm this as well?

The other thing is that these trials all exclude people within the first three days who didn't get enough drug. So there is very little information on what happens to people in the first three days in modern trials.

So comparing the 91 percent success rate from a current trial to 1.3 percent success rate 70 years ago violates every part of the constancy assumption because we're comparing an endpoint that's out way beyond the end of therapy to a three day outcome, and every one of these trials says in the per protocol analysis, you have to get at least three days to be evaluable.

So, all those failures in the front, they're called indeterminate and taken out of those per protocol analyses.

So, before we get too far down the line saying how great an idea is, there's a lot of devils in the details here about how to actually analyze all of this.

DR. ECHOLS: I said there was -- needed license in terms of interpreting the data --

DR. POWERS: Poetic or scientific?

DR. ECHOLS: Well, but the point is, I mean, ICH guidelines -- and you said earlier in your talk, it's -- because there is no evidence of treatment effect, therefore it is ethical to do a placebo controlled trial. I don't accept that either, because the little matter of what's safe for a patient and there are -- as I tried to point, even for sinusitis, which no one would suggest has a major morbidity outcome measure if you're not a treatment success immediately, but there is still -- patients have pain. Patients have other aspects, particularly when you throw in, "I want to stick a needle in your sinus, but

I'm not going to give you any antibiotic," that's why IRBs and ministries of health are rejecting, even these simple disease conditions and absolutely categorically, even when they know there's no evidence, proof that antibiotics work, they still consider it unethical to conduct the placebo controlled trials.

DR. GILBERT: Roger, you've definitely gotten our attention. Now, I've got patients that -- questioners on the floor have been very, very calm about waiting here. If you can ask real quick, your questions, we'll give you time.

MR. NUSRAT: I'm Roomi Nusrat from Sanofi Aventis, and for the record, I'm a physician trained in both and certified in infectious diseases and pulmonary diseases. So this is all close to my heart.

First of all, I was not going to say this, but, John, the MITT population includes patients in the -- that have not

received the first treatment and as Karen can sometimes -- you know, at a later time, we can discuss this.

Those patients are included in her presentations, and I know our data from Sanofi Aventis, so --

DR. GILBERT: You and John will have to discuss that outside, and we really have to move on. So ask your question real quick.

MR. NUSRAT: So here is the question. Henry Masur is not here today. But he once said to me, as I was struggling with the current issues, that it is you in the industry that has to develop drugs. We don't.

So that's the starting premise. We have to work with specific guidelines, and we have to have reasonable outcome measures to target, and at the end of the day, we have to enroll patients, and I think Roger has articulated that perspective, and I think it's probably understood by most people in the room, that that challenge has to be -- that

challenge has to overcome.

I don't know if anybody if -- when you were younger, read Fisher and Ury's book, "Getting to Yes." I think that the needs of the scientific community, all of you, including myself, is to do good science.

At the same time, us, the other side, the Darth Vaders, the industry, we need specific guidelines and what's not only suffering is the patients with mild sinus disease, is the patients with resistant tuberculosis, malaria, the patients in the intensive care unit.

I think that what we would like to ask you to do is, we have to come up with some interim guidelines, so that we can target -- so that innovation can continue.

DR. GILBERT: We agree with you 100 percent. That's why we're here. How long is your question, ma'am?

DR. KAMICKER: Mine is just a comment. I'm Barb Kamicker from Pfizer. I'm

a microbiologist, and I'm addressing Dr. Fleming's comment about how good these antibiotics are.

I ran infectious disease models in rodents for 15 years, and believe me, you are not going to advance a compound unless it really looks good, whether it looks good against your non-infected control and whether it looks good against the comparator. It has to be at least as good against a comparator before it's going to advance.

So, I find nothing astonishing that -- about these state of the art compounds that look 93 percent efficacious.

DR. FLEMING: But that didn't address my issue though. Clearly, you're going to advance something that you see has potential benefit. We do that across diseases all the time, to see that everything -- to claim that everything that you're advancing has essentially the same benefit is quite implausible, but that's only part of the

issue.

Dr. Powers was bringing up the other issue. We're basing all of this issue on slide 18, and slide 18 is clearly taking serious liberties in terms of really having reliable comparative evidence, and it gets you to a margin of about 15 percent, although there are some re-calculations we could do as to whether it's even that.

But it really comes down to then is assessing the degree to which we can reliably say that that 35 percent in today's world, in today's trials, would have had essentially no response.

DR. GILBERT: Okay, we have to cut this off, although I'm hesitant to do so, because it's getting right to the guts of the issue.

We're going to go to the flip-slide now, which is equally important, which is safety, and I think Tom is going to introduce the next speaker.

DR. FLEMING: Thanks, Dave. So, as Dave says, a key aspect of benefit to risk is assessing safety, and we have a few presentations to bring us through some of these key issues and challenges in safety.

Our first speaker is Bruce Psaty, who is Professor of Medicine and Epidemiology and Health Services at University of Washington.

DR. PSATY: Thank you, Tom. Thank you to the organizers for inviting me here today. I have several disclosures, and I come with kind of a split past, a divided past.

I'm a general internist. I practice at the county hospital in Seattle, and I'm also a cardiovascular disease epidemiologist with experience and expertise in study design and drug safety, but I don't come with a history of having done studies in the setting of infectious diseases.

The general argument I'm going to make today is that high quality evaluations of

antibiotics are essential to characterize the risk benefit profile and that inadequate evaluations, actually of either side, of efficacy or safety, compromise the knowledge base for physicians and for patients.

Now, as an internist, I was struck in looking at the early trials in pneumonia. It actually is a situation in which the historical controls work well for the septicemia with pneumococcus, 80 percent mortality. We had a nice slide on that earlier, and after the introduction of penicillin, it's down to about 20 percent. Not that much improvement for the serious septic patients since then.

In world historical terms, we've got what I would characterize as an epidemic of antibiotic use, and this is the result of physicians who trained, like I did, were trained to think about treating infections aggressively and to using antibiotics to kill the bug, as opposed to perhaps improving the

outcomes for the patients.

Not long ago, I think there were arguments against -- and there still are apparently, using placebo controls for acute bacterial sinusitis. But I have to say, in our clinic several weeks ago, we discussed the Williamson article looking at a placebo controlled trial in antibiotics, and that is actually a very important trial.

It's important because it will help eliminate the use of antibiotics where there's little or no benefit and where there's only risk, there's only risk, and that is very important.

I have to confess additionally that I sit on the events committee for the cardiovascular health study, cohort study for 5,888 older adults, and I see how antibiotics are used in four different communities, and it's not unusual for a little old man to come in to the hospital and to have a funny looking chest x-ray and to be short of breath and you

get a little Lasix and a little antibiotic, and this is not necessarily optimal therapy.

As an epidemiologist, I really -- I was innocent and naive, I actually looked for the placebo controlled trials, to see what the anchor is, and I'm glad to see that there aren't any and I didn't fail to miss them.

I'm concerned about the un-interpretability occasionally of findings, where there aren't good anchors.

I looked at the community acquired pneumonia guidance from the FDA. I think I looked at the most recent version from July 1998. As an epidemiologist, I was concerned about the failure to assist on ITT analysis. The use of evaluable patients, when you exclude those who stop therapy, who die because -- and the cause of death is attributed something other than pneumonia, this breaks the randomization and turns what is a high quality trial into, potentially, an observational study.

There is also a failure to insist on double-blinding, although I noticed in Dr. Higgins's talk that the standard apparently for the trials that are coming through is that they be double-blinded.

I've mentioned the issue with the non-inferiority design. There is no anchor, and then you need high quality data. People have mentioned earlier today that noisy data contribute to a finding of non-inferiority.

There is potentially a bias with using only the evaluable studies, and I looked at several meta-analyses, and in the meta-analysis by Salkind -- I think others have referred to this today, the intention to treat analysis had a different finding from the evaluable analysis and these are odds ratios in a meta-analysis for cure rates, and the bias can actually represent 10 to 30 percent of a typical non-inferiority margin in these trials.

There is also an open trial bias,

and in a meta-analysis of the quinolones versus -- or macrolides versus the beta-lactams, they looked at studies where there was concealment of randomization versus where there was not, and here, there's a relative risk. It's been switched, in terms of the direction of effect, and the studies with unclear or inadequate randomization tended to show a higher, much larger benefit than those with adequate blinding, and the bias here represents probably 25 to 50 percent of typical non-inferiority margins.

And then I also wondered about whether placebo would be ethical. There's a review of clarithromycin, which remarks that the cure rates over the last 10 years with the drug have remained remarkably stable, even though resistance to the drug has increased from five percent to 25 percent or so, and it occurred to me that there might actually be an alternative explanation. In fact, maybe the drugs are not doing much, whether or not the

bug is resistant, and I pulled data out of the meta-analysis that has been referred to earlier, in which -- which Dr. File referred to earlier, and beta-lactams really are functionally placebos here in patients with Mycoplasma or Chlamydia, and at the point of the test of cure, there's really no difference.

Now, he did point to several studies where symptoms may have resolved sooner, but if there is an opportunity for a placebo controlled trial -- and placebo controlled trials are potentially important from the point of view of public health, because they tell us when we might not need to know. A placebo controlled trial tells you does a drug work. Is there any improvement? An active controlled trial tells you which is better.

Knowing whether the drug works can identify situations, from the point of view of public health, where it's not needed. If we

can come up with some of the PCR tests to detect pneumonias and types of pneumonia early, it would be very valuable to know if there are several types of pneumonias for which -- bacterial pneumonias for which we don't need treatment.

To step back for a minute, I worked on the IOM Drug Safety Committee, and we thought about drug safety not only in the pre-approval but the post-approval setting and thought about assessing safety throughout the life time of a drug.

There are many withdrawals that occur after drugs come on the market. In one review from `69 to 2002, about 75 drugs were removed from the market, 11 with special requirements that are effectively removed from the market. In another, for 584 new chemical entities, 45 received black box warnings, and 16 were withdrawn.

So the information that we acquire about safety isn't always present, and in

fact, it looks like in about 10 percent of the drugs, we get significant new information after the drug is on the market, and this is part of the normal course, I think, of drug safety.

Many of you know this process better than I do. There is pre-clinical information to assess toxicity and for several of the antibiotics, this turned out to be quite valuable. For sparfloxacin, that's where the initial change in QT was detected. For telithromycin, there was liver toxicity noted in rats, as well as other animal species, and we have a series of studies to evaluate the drug for approval, and then in the post-marketing setting -- there were various studies, and an Adverse Event Reporting System that is especially weak for detecting adverse events, except for severe and rare ones that are completely unrelated to the indication for the drug.

So we don't have a very sensitive

system for detecting adverse effects in the post-marketing setting.

There is considerable asymmetry in terms of safety and efficacy during the evaluation process. For the phase three trials, they are designed and powered properly, and approval is contingent on evidence about a non-inferiority margin, about an effect.

The safety evaluation is always more ad hoc than that, and the FDA guidance on the pre-market risk assessment is really quite good in pointing this out. The adverse event data are collected, and it really becomes a kind of diagnostic act to notice and define an emerging safety signal.

Based on adverse events alone, there were 25 drugs removed between 1978 and 2003. A number of these were antibiotics, the quinolones figure heavily here, and for a variety of severe or potentially severe and serious adverse effects, hemolytic syndrome,

long-QT arrhythmias, hepatotoxicity, phototoxicity and hypoglycemia as well.

Many of you know more about these drugs than I do. There are common side effects, the GI and CNS side effects. There are uncommon ones that are potentially serious, and if we're using these drugs for patients who are not likely to receive much benefit, the -- our tolerance for safety issues has to be less. We really need for the individual patient to be assured that the risks will not exceed the benefits for treatment.

This is a report of a study looking at the IC50 for the HERG potassium channel, which is the primary mechanism by which these drugs prolong QT, and there is a range of sensitivities with sparfloxacin, comparing its IC50 to the peak-plasma level, to having levels that are quite close to 10. Grepafloxacin with 16, moxifloxacin with 22, and there was a subsequent study looking at

the safety and efficacy of moxifloxacin versus levofloxacin, and it included 394 hospitalized patients greater than 65 years old with community acquired pneumonia. It excluded the severely ill, and there were only 71 percent who were evaluable.

The cure rates were comparable, 93 percent for moxifloxacin, 88 percent for levofloxacin, with a 95 percent confidence interval for that difference of -2 to +12, and in the safety study, they concluded that cardiac rhythm safety was similar.

Well, the data from that study come from Morganroth's paper, in which there was a composite outcome about ventricular tachycardia, as well as sudden death, and this shows the counts of events and the relative risks, and we have a relative risk for the composite endpoint of 1.6. The 95 percent confidence interval goes from .3 to 3.5, and for death during treatment, there was a two-fold increase.

Now, admittedly, the confidence interval is quite wide. It's .5 to 8, and you know, I guess, in a very serious condition, where the mortality is high, one would be willing to tolerate a large increase where there was a clear benefit for this particular therapy, where the risk of death could be as high up as two, four, five, but in patients who are -- otherwise have mild conditions that may resolve on their own, we don't want to expose those patients to drugs that may have this sort of toxicity.

I think I would not, myself, conclude that this study shows cardiac rhythm safety. I would conclude that this study actually gives you an estimate of what the effect size might be if you did a larger trial that you would want to detect.

So this study is so small that it doesn't actually provide a lot of confidence about the cardiac safety of this drug, and it really confirms the potential signal for the

difference between moxifloxacin and levofloxacin.

The sponsor apparently felt they needed additional study, and they conducted a clinical experience study, and this was published in 2004. It's relatively recent. Eighteen-thousand patients received the drug for five to 10 days, and the indications included mild to moderate pneumonia. Astonishingly, the patients were all enrolled within about two and a half months, and they report 900 C- 297 cardiac events, they had ECGs on 122.

It turns out, this study had no control group and, I think, provides no useful information about the safety of this drug in clinical practice. This looks to me to be a seeding study of the sort that Bob Temple wrote about back in the New England Journal in 1994.

For telithromycin, there was another safety study conducted in 24,000

patients. It was able to detect no difference in hepatic adverse effects. The data submitted included data that were suspect and fraudulent, and this large study was incapable of detecting liver -- adverse liver events.

On the other hand, our insensitive marker, the post-marketing AERS data, identified a rate of acute liver failure of 167 per million person years, which was 10 times the rate for levofloxacin.

Now, admittedly, this is a rare but serious risk, and it's not a risk that would be tolerable if the benefits for the drug are small to minimal.

From the point of view of public health, the use of antibiotics in situations that are not helping patients contributes to drug resistance. In the Netherlands, antibiotic use is about a third of that in France. Penicillin use is about 40 percent, and the rates of pen-resistant Strep pneumo are remarkably different, and, indeed, the

cross-national correlation between rates of antibiotic use and drug resistance are extraordinarily high, .84.

So some concluding observations. I think there are opportunities for improving the study design. I tend to prefer ITT analyses. This actually provides us with denominators for risk and benefit within the trial, if they use the same people and the same numbers of people. Blinding looks like it's already being done.

I think that we have an obligation to provide the optimal therapy as comparator with -- when there are known benefits. I favor mortality as an outcome in severe community acquired pneumonia. I think that there needs to be an improvement in the safety evaluation, and this means in part -- I think the best opportunity is for safety, other than common adverse effects that you're likely to see and in fact, I see probably more antibiotic associated diarrhea in my clinic

than I do see pneumonia.

In order to do that, I think the safety evaluations need to identify signals, and I think the FDA is very good at this, and then follow them with high quality studies, not clinical experience studies, and not small, underpowered studies.

I think it's reasonable to consider DSMBs for many of these trials, and as a clinician, I would be reluctant -- I've seen patients with galloping strep pneumo infections. I'd be reluctant to randomize those patients to placebo, but there may be other conditions, including the Chlamydia and the Mycoplasma, where placebo trials have a role.

So, I thank you, and I'd take any questions or comments.

DR. FLEMING: I think we'll do the questions together. I think what we might do is do the questions together on the safety presentations. Ed?

DR. COX: Next, I'd like to invite up Tatiana Oussova, who is a Medical Officer in the Division of Anti-Infective and Ophthalmology Products at CDER, FDA, and Tatiana is going to be talking about evaluation of drug safety in community acquired pneumonia.

DR. OUSSOVA: Thank you and good afternoon, everyone. In today's presentation, I'm going to concentrate on the pre-marketing assessment of drug safety in community acquired pneumonia, and this is just a brief overview on our approach to drug safety. This is my disclaimer. I hold no financial conflicts.

As a regulatory requirement for approval, a drug needs to demonstrate sufficient evidence of efficacy and safety, and as the Food, Drug and Cosmetic Act states, any new drug application should include all tests reasonably applicable to show the drug is safe and, this is important, on the

proposed labeling, and the results of such tests should show the drug is safe under such conditions.

Safety assessment of a drug begins at the very early stage in drug development and safety data are continuously evaluated throughout the stages. It starts with non-clinical data that identify target organs of toxicity and determine therapeutic dose safety margins for future clinical trials.

Then it comes data from phase I and II clinical trials that predict possible adverse events in phase three trials. It also allows for design safety assessment for phase III trials, that is to tailor safety monitoring to anticipate its specific adverse events in phase III trials.

However, due to limited exposure in phase I and II trials -- and we are talking about few hundreds of patients, serious adverse events are rarely identified.

When a new drug application is

submitted, the goal of its safety review is to critically examine the sponsor's contention that their drug is safe for its intended use, and what does it mean? It means that we assess whether the testing for safety was adequate. We determine how significant the identified adverse events were and how they would impact on drug approvability.

We describe the safety issues that should be included into product labeling, and we decide whether additional safety studies would be needed.

What are the data sources that are reviewed? It includes randomized controlled trials, open label trials, post-marketing experience, if there is such, and it could be foreign data if the drug is already marketed outside the United States or even if it's marketed in the United States for different indications.

It includes medical literature, and we would consider a safety profile of other

drugs in the same class, even if approved for other indications.

We take the following approach when we review the NDA safety data base. We would first characterize the population based on age, gender, underlying medical conditions, and other factors that may influence the outcome of the study.

We would characterize the dose and extent of exposure. We will identify adverse events and then assess the relationship between the drug and the adverse event, and we try to identify the risk factors for serious adverse events and for those adverse events that are common in general population, it is helpful to look at those events rate in a comparator arm.

What do we want to know about exposure? When we characterize the magnitude of exposure, we want to know whether there was an adequate exposure in terms of the number of patients and the duration of treatment at the

intended dose, and if the labeling recommends a dose range, we would like to know how many patients were exposed to the highest recommended dose. We also want to know were there any special populations, such as renally or hepatically impaired included into the study.

When it comes to assessing adverse events, the following are the most concerning to us, death, serious adverse events, and discontinuations due to adverse events. When we are looking into these adverse events, we always assess the causality, that is trying to answer what is the likelihood that the drug had caused those adverse events.

Other important parts of the safety review are common adverse events, laboratory data, vital signs data, ECG data, and safety in pregnant women and special populations, such as elderly or renally impaired.

These are specific safety issues that we usually address with antibiotics,

liver toxicity, renal toxicity, allergy related toxicities, cardiac repolarization, or QT studies, however, those are not unique to CAP, and they are common across other drugs as well.

Despite the robustness of data submitted with a new drug application, there are inherited limitations to what we can learn from the NDA safety database, and there are several reasons for this. One is -- we always deal with a limited exposure and this is about just a few thousand patients included with the NDA safety database, and therefore, rare serious adverse events are not usually captured, and when I'm talking about this rare adverse events, I'm talking about adverse events that occur in order of one per 10,000 or 100,000 patients.

However, observing no adverse events should not be interpreted if there are -- as there are no risks, and it simply could be not -- just unknown at the time of NDA

review.

Other thing is that studies are not designed to address specific safety questions. They are powered for efficacy, and they have no pre-specified safety endpoints.

Other thing is -- and this is particularly true for sick patients in intensive care settings, it is very difficult to ascertain serious adverse events in this sick population. Sometimes, adverse events are erroneously attributed to underlying disease or vice versa to the drug.

The NDA review results in either approval or non-approval of a drug. After we complete our review of efficacy and safety, we perform risk benefit assessment and we make a final decision, and if we have any questions about risk benefit assessment or specific safety concerns, we can always ask for input from an advisory committee.

When a drug gets approved, the results of the safety review are applied to

the product labeling, including patient education, education materials, and we also develop a surveillance plan to further evaluate non-serious risks and identify unknown potential risks.

Assessment of drug safety does not end after the NDA gets approved. The sponsor continues to monitor for adverse events, and they submit periodic safety updates and annual reports. There is also Adverse Events Reporting System or MedWatch, which is voluntary system where anyone can report adverse events associated with a particular drug.

As the result of post-marketing safety findings, the labeling changes and updates occur, and usually, they occur in adverse reaction section where we include post-marketing adverse events reports or warning section and with possible elevation to a box warning or medication guide.

This is basically the end, and to

conclude my talk, I just wanted to mentioned that this pre-marketing safety evaluation, as I just described, is not unique to CAP indication, and it's applicable to all drugs across all divisions of FDA.

DR. COX: Thanks, Tatiana. We'll continue to hold questions until after Dr. Talbot's presentation.

DR. GILBERT: So I'm pleased to introduce George Talbot, independent consultant to industry, also pleased that George has been an invaluable member of the task force of the Infectious Disease Society of America on the availability of antimicrobial agents. Thank you, George.

DR. TALBOT: Well, I have to say good evening, everybody. It's no longer good afternoon, it seems like, and thank you to Dave, Tom and Ed, for asking me to speak and also awarding me the coveted last speaker of the day award or position.

In fact, I noted, they had to give

me a page three to just fit my name on. So, hopefully it will be worthwhile your having waited.

So here is my assigned topic, again, thank you to Dave, industry experience and importance in monitoring safety.

Like some other speakers, I had some difficulty with the title, and I accordingly made some qualifications to the assigned title, and they are shown on this slide.

First of all, contrary to rumor or statements on the agenda, this presentation is not an industry perspective, because I don't really know exactly what industry is or what industry thinks about this topic. So I'm only presenting my thoughts on this and really not anybody else's.

I also knew that there were two speakers ahead of me this afternoon or this evening, and I hoped to have something left to say, other than "she said it." So I took a

somewhat different tack, and I'm going to focus on some different topics and perspectives that hopefully will be useful to the audience. My goal is to stimulate discussion.

The discussion points are shown on this slide and I thought it would be interesting to discuss the non-safety facets of safety. That's what I've called them. We could probably come up with a better name, but that's what I could think of a few days ago.

Really, the point I'd like to make is that efficacy considerations in designing a clinical trial are just as much about patient safety as "safety" is. The two prior speakers spoke about safety in a classical sense. I'd like to leave you with a thought that efficacy components of study design and implementation are really about safety as well.

And so, I'll discuss where can we go wrong in this aspect and where can we thereby put our patients at risk, when they

participate in clinical trials. I'd also like to spend a minute or two to discuss approaches to mitigating the safety risk of efficacy.

I will spend some time on some somewhat random thoughts on traditional safety issues, such as those discussed by the prior speakers, and then I have some conclusions.

My disclosures for the CAP workshop are shown on this slide. I was recently Chief Medical Officer at Cerexa. That ended in October and currently, I've resumed consulting to industry and the most relevant potential conflict, I should disclose, is that I still consult for Cerexa, which has an ongoing CAP program.

So, what are some of these non-safety facets of safety? Well, as I mentioned already, efficacy is really another facet of safety. This is not a surprise, but in fact, we often speak in an efficacy safety dichotomy, and so, my observation is -- and it's based on a broad experience over 15 or 20

years, is that when we're talking about efficacy, we're not necessarily always thinking so much about what the implications might be for the safety of patients participating in the study.

This happens because there's a press of other clinical development considerations. Monitoring efficacy during studies is time consuming, expensive and in particular, is constrained statistically, and I'll come back to some of these constraints a bit later, and I think it also relates, perhaps, to an overly narrow perspective about what constitutes safety in the clinical trial process.

Sometimes, unfortunately, there can be a tendency to forget that there is a patient at the end of each clinical trial protocol, not often, but it's something we need to continuously remind ourselves of.

So, what can go wrong from an efficacy perspective? Well, the first thing

I've listed here is dose selection. The rest of my list includes dose selection, and then, I would change the pace, two other things. One is the choice of comparator, the choice of adjunctive anti-microbial therapy, plus the impact of prior anti-microbial therapy, not so much on the integrity or results of a study itself, but on what might happen when the compound gets into the market place.

I think another thing I'd like to highlight is sub-optimal adjunctive non-anti-microbial therapy.

So, what about dose selection rationale? We begin now -- and I think there are very good points made earlier about the sophistication of dose selection, but we usually deal with these usual suspects, as I've put there -- let's see if they're there.

So, we start with the in-vitro-data, we talk about or evaluate efficacy in animal pneumonia models, PK data, we use known PD relationships in plasma, do our modeling,

to some extent, evaluate lung penetration of the compound and consider the active site of the drug in the lung and we also integrate into this approach, the prior experience with the class, as well as phase two data, if they are available.

Now, what can go wrong? There are some unexpected events that can happen. First of all, in any study, and this is not so much just CAP, but it could be HAP or intra-abdominal infection. Sometimes, the spectrum of organisms can be different than that, that was anticipated.

We could mention -- I'll mention one in a moment, but if you picked your study drug because of its spectrum of activity for a certain group of organisms and you encounter a different spectrum of organisms, that's a problem.

Another event that could occur is that if the target pathogens are those that you expect, but their MIC's to your study drug

are higher than you would expect. That's going to decrease the achievable PK/PD index, potentially with deleterious consequences.

Related to activation -- pardon me, to activity at the target site, is the possibility of drug inactivation at the target site, something that we don't necessarily think about much, or didn't, but certainly has, in at least one instance, proven to be a problem for the safety of patients participating in a study.

Unanticipated PK variability can be a problem and unanticipated drug/drug interactions could also be a problem.

So, let me give you some specific examples. Drug inactivation of the target site, daptomycin, which failed -- and I think Bob Arbeit mentioned this earlier, failed in its CAP program for very unexpected, unanticipatable event and that was inactivation by surfactant and I would mention that I would give kudos to Cubist for

publishing those data in JID in 2005. They thoroughly explored the reason for this failure, and that's benefitted other companies who have followed in this field.

And here we see the explanation for that. This is from the abstract. What was interesting was that there was efficacy in two animal models, but not in a third, Strep pneumo and simple bronchial-alveolar pneumonia. That's the sort of signal that can really, in my experience, be difficult to pick up and move with, and in retrospect, it tied in neatly with what was being seen, but unfortunately in these studies, this effect did become evident and the hypothesis then was complete. So, we do need to consider that.

Since that experience with daptomycin and Cubist's discussion of it, sponsors have responded, and I think appropriately so. Pulmonary PK studies have been performed for novel compounds. I think that's particularly important when you're

dealing with novel compounds with no recent experience, as opposed to classes with a known effect, and we see it was done for tigecycline, telavancin, iclaprim and oritavancin.

In addition, telavancin performed a surfactant interaction study, which I think also undoubtedly added a great deal of confidence to the data they had obtained from animal pneumonia models specific for their target pathogens, specifically MRSA, and so, there was really a consistent, and I think impressive attempt and effort to identify the potential safety risk of efficacy in their upcoming clinical studies.

Now, what about different organisms then anticipated? We usually think we know very well, what the spectrum of disease is, but one potential problem on the horizon is community associated MRSA and thinking about design of studies for CAP.

This bug is still rare, especially

in a clinical trial setting, as a cause of CAP, extremely rare. My personal opinion is that although these studies should apply relevant exclusion criteria to eliminate subjects at risk of MRSA, that it's not yet necessary to include MRSA coverage in trials of CAP. But we must certainly be vigilant, as to when MRSA coverage should become routine.

A related issue here is that if you have a known set of pathogens, but they have expected -- pardon me, MIC's higher than expected, that can also be a problem.

A third example I'll give you, and this will be the last, is unanticipated PK variability, resulting in sub-optimal exposure, and I have a couple of possible examples with apologies to George and Paul -- that's not Paul McCartney and George Harrison. That's George Drusano and Paul Ambrose, who have taught me a lot over the years.

A possible example could be that the target population you're studying in your

trial, differs from that studied previously and this new population has a higher clearance or some other parameter that could result in decreased drug exposure. Same would apply for drug/drug interaction.

Now, a possible example of sub-optimal dosing, this was in HAP, but not CAP, was in the study of HAP reported by Wyeth, and again, kudos to them for publication where there was success in CAP, but a failure in that sub-set in HAP.

So, this, I think, shows that this is more than a theoretical concern and it happened despite the fact that Wyeth did vet their dose selection rationale extensively and in fact, had conducted a pulmonary PK study to assist in dose selection.

Moving from the dose selection rationale, I mention comparator in adjunctive therapy. There is the ICH guidance on comparator therapy, but it's critical for the safety of our patients not to include "straw-

men", maybe it should be straw-people now, straw-people comparators. They have to be given an appropriate dose and dose frequency, which in fact, may have changed since the initial regulatory approval, as the epidemiology of bugs have changed. The comparator has to have an appropriate spectrum and an appropriate tolerability profile, so that you're really giving the patients a fair shake at an optimal outcome.

Adjunctive anti-microbial therapy is also problematic in some respects. Particularly, if the spectrum of the study drug is not broad enough for all likely pathogens, adjunctive therapy will be necessary. Optimal adjunctive therapy should be employed to ensure the best overall outcome for both treatment groups.

This may be more relevant in HAP, for example, but it can be true in CAP as well and it does highlight, I guess, one hazard of NI trials, which is if your adjunctive therapy

is not optimal, probably both treatment groups will have lower response rates, but it won't necessarily affect your finding of non-inferiority.

A specific conundrum in CAP was alluded to earlier, but it's illustrated on this slide. What if the spectrum study drug does not include atypical pathogens? How do we provide optimal therapy for patients without overlapping coverage, that confounds interpretation of efficacy?

Studies of cephalosporin therapy for CAP are really right there, especially in the U.S. It's difficult to enroll patients now, without adjunctive macrolide therapy. That's something that will be discussed tomorrow, but it obviously represents a hurdle for design of clinical trials and conduct of clinical trials and it's an important question that warrants further discussion.

Now, prior therapy can also have an impact on the safety aspect of efficacy. This

was another dapto. experience, which again, I congratulate them for publishing, and what they showed is that prior effective therapy, in some cases less than 24 hours of prior antibiotics did have an effect on efficacy.

So, we knew that could happen. It's been published. The problem is that this aspect of safety for patients really becomes later -- apparent later, only post-marketing, when the drug may be used without the benefit of prior anti-microbial therapy.

The solution for clinical trial design, which is to avoid all prior anti-microbial use, poses major logistical consequences and difficulties and we clearly need some better approaches to this issue.

Now, what about adjunctive non-anti-microbial therapy? Clearly, for our patients, outcome can be compromised by inadequate adjunctive therapy, but this is more obvious for surgical diseases, such as complicated intra-abdominal infection. It's

less obvious for non-surgical conditions.

What I would submit to you for your consideration is that for CAP, we do need to consider how to optimize adjunctive therapy and not ignore that, again, under the guise of the -- or the protection of a non-inferiority design.

So, we need to avoid poor pulmonary toilet, sub-optimal respiratory therapy support, inadequate mobilization of patients, if they are in the hospital and premature hospital discharge, among others.

So, how can we mitigate the safety risk posed by problems with design related to efficacy? First of all, rigorous attention to dose selection, prior to phase two. The dose selection should be thoroughly vetted with external people with expertise in that area, to make sure that there are no holes in your argument, and I also urge you to make use of the FDA end of phase two meeting, where a full dose selection rationale has to be articulated

and is a very useful exercise, and I've mentioned the importance of these last three bullets.

So, some final thoughts on this part of my talk. I think we have an obligation to consider efficacy as a safety issue that extends beyond the clinical trial period. It's imperative to reflect efficacy issues that impact patient safety in the product label, and I think an excellent example is again, what Cubist did, with relation to their labeling for their first approval. They included the words `Cubicin is not indicated for the treatment of pneumonia'. Post-marketing risk minimization programs should consider this aspect of safety.

Now, a few slides and I'll be done, to hopefully keep us on schedule. Just a few comments on some selected traditional safety issues. The following four bullets highlight the points I'll make briefly, related to FDA guidance documents, internal safety assessment

processes, DMC's mentioned by a previous speaker and an approach to infrequent events and possible signals.

I'd mention that FDA has articulated and published some excellent guidance documents, which are particularly useful for smaller companies who may just be starting in this area, related to adverse event reporting, development and use of risk minimization action plans and pharmacovigilance practices, among others.

Associated with this is the need to have clear internal safety assessment processes. I think this is less consideration for larger companies that have a long history, but for start-ups and new companies, it's critical to have an a priori defined safety assessment process that will ensure the safety of patients in the study or studies, and this is not just a question of meeting the regulatory literal requirements. It's not just reviewing SAE reports, and completing the

reporting.

What's really needed is an ongoing attention to the big picture of the emerging safety profile and I can't emphasize that enough.

Another piece of advice I'd give to companies in that situation is not to wait for a problem to appear to establish a process. You have to have a process identified a priori, so that potential signals can be evaluated promptly, using a multi-disciplinary approach and part of this is to be considering the advisability of seeking external expertise and the objectivity associated with that, at some point during your assessment of any changes or unanticipated findings in the AE profile.

Now, what about rare events and possible signals, also a point mentioned by a previous speaker. Easy to say, but in my experience, it's one of the most difficult aspects of responsible safety monitoring with

the patient in mind.

As we all know, early on, you have small numerators and denominators and in that setting, it's very difficult to remain free of bias, even when you acknowledge to yourself that you could be biased. That's one reason to use external expertise to help you with that and it's a reason to keep an open mind, so that you avoid constrained hypotheses, and one must also look at these cases' possible signals in extreme detail and consider a vast array of potential explanations, other than the baggage you bring with you a priori.

This guidance from FDA, again, is very useful in this regard and I would mention that a lot of the points in there are extremely worthwhile to keep in mind.

One of things I see is that people feel there are obstacles to use of DMC's and there are some, and I've listed them on this slide. It takes time, effort and considerable expense to establish DMC's, at a time when

you're trying to move things along rapidly.

It can be difficult to find the right people to sit on them, who have the expertise and aren't part of your investigational staff or otherwise involved, competitors, etcetera.

There is concern about maintaining the integrity of your clinical trial. There's also the operational concern about getting data to the DMC in a timely fashion, so that relevant decisions can be made. If your complicated skin study is enrolling in nine months and you want a mid-point analysis, you may have your data about the time you're finishing enrollment.

There's also a general concern about loss of control, when one establishes a DMC and also, I think, not thinking of efficacy as a safety consideration.

Now, the advantages to use of DMC's, I think, are exactly parallel to the potential obstacles. One is that with the

proper use of DMC's, you may, in the end, save time, effort and expense, in many ways. It's also better to find the right expertise to help you with your decision making sooner, rather than later.

I also believe that DMC's can actually be constructed, and the guidance makes this point, to ensure that the integrity of the trial is not only maintained, but perhaps improved.

The constitution of a DMC also ensures that you will be working hard to ensure timely access to data, so that relevant decisions can be made in the patient's best interest, and I think in some senses, DMC's actually give you improved control over your study, as opposed to the fear of loss of control.

Finally, consistent with my hypotheses in this discussion, I think that having a DMC, in selected situations, not all I don't think, but in the right situations,

highlights efficacy as a primary safety consideration.

So, in conclusion I would say that efficacy must be considered a patient safety issue. I think the analytic approach that was described a few minutes ago, to evaluating your safety database, collecting the data, it's something we all know and it's obvious that it must be done, but we must also consider efficacy as a patient safety issue.

I think that steps can and should be taken during the planning and execution of clinical trials, to ensure that optimal efficacy is achieved and that it does not become an unexpected safety issue.

Finally, smaller companies must take time to develop a process, as I described a few minutes ago.

So, my final thoughts are, don't cut corners on efficacy risk minimization. Remember, there is a patient at the end of every clinical trial protocol, and overall,

keeping this in mind is going to be not only in the patient's best interest, it's going to be in the best interest of your drug and your company. Thank you for your attention.

DR. GILBERT: Thank you, George. If you want to stay up there, we have time for maybe just one or two comments or questions on the safety presentation. Yes, Barry?

DR. EISENSTEIN: Just a brief comment. George, very nice overview. I'd just like to add something to the Arbeit data, with the lack of efficacy of daptomycin in CAP, and we're going to hear tomorrow, from Paul Ambrose, a little bit more about that data.

But to talk about two of those things, one was the prior effect of antibiotic therapy. That has a major effect, obviously, on being able to see the Cubicin effect versus the comparator and does raise, as you say, the major issue about how can you do these sort of studies in the United States.

But what's also interesting is that if you look at those who did not get prior effective therapy and then view Cubicin as presumably, no worse than placebo, you could set yourself a floor for a placebo comparison, because there is a clear-cut superiority signal that ceftriaxone has over Cubicin, so you at least, in contemporary time, have a therapeutic effect.

DR. TALBOT: That's a good point. Well, those data will be discussed further tomorrow. Yes, I was -- 65 percent versus 90 percent or something, 75 versus 90, yes, right.

So, I think again, that does support a treatment effect in your patient population.

DR. GILBERT: Okay, Robert has a question.

DR. O'NEILL: Yes, I have a question to Dr. Psaty and to you. It relates to a conversation that we're going to have on the

design of studies and particularly, in the mild CAP area.

It relates to the issue of, one of the things you didn't put down, in terms of where can you go wrong, and I think it's a critical component of what's hard about this area, and it's misdiagnosis.

When you think you're treating the disease that you are, but you're not, and it's part of the inclusion in the current clinical trials -- and you made the point that if you can't benefit from the drug, but you share all the risk, that's a real problem.

So, if you have a drug that essentially has a serious risk profile and you're giving it in a mild condition and you essentially don't take the pains to make sure that the entrance criteria rules out those folks who aren't going to benefit, what's your comments on that and what are your thoughts, in terms of a fix, because that's where you are, in the mild area.

DR. TALBOT: Should I start? Well, first of all, I agree totally with your point. It hadn't occurred to me, as I was putting this together, but I think it's very germane.

Just as we have a "safety obligation" for patients who are enrolled, we have exactly what you say, we have an obligation to enroll patients who are appropriate to enroll and who could benefit from the treatment. So, I agree with you.

In terms of how to improve that, I think that there are two aspects. One is what was discussed earlier, about mechanisms to enrich trials for the pathogens that we want and need and that presumably cause disease, and the other is in the conduct of the trials, in terms of how those patients are managed.

DR. GILBERT: Thank you, George. In the interest of time and as George was saying, tongue in cheek, we're entering the cocktail hour. We do want to go around the table now, and I'll ask Ed and Tom, after I make a

comment, as to what we would like our panel members to address, and we have three points written out, actually, in the program. I don't think I have to read them out loud.

In short, you've heard a lot of data and we -- I'm serving as the Assistant Rapporteur with Brad, and we just want to get everybody's viewpoints out on the table, with respect to design, superiority, placebo controlled versus non-inferiority and endpoints are really the two major points. But feel free to bring up anything else that's on your mind.

This is your chance to air whatever issues are pivotal in the construct of clinical trials for mild community acquired pneumonia. We'll have a similar session at the end of tomorrow, on the more severe hospitalized patient with community acquired pneumonia.

DR. ECHOLS: David, can I just get a clarification on the second bullet?

DR. GILBERT: Sure.

DR. ECHOLS: You say that -- how likely is it that superiority could be demonstrated in a controlled clinical trial. Are you referring to active control clinical trial, placebo control clinical trial?

DR. GILBERT: Either.

DR. ECHOLS: Okay.

DR. GILBERT: Depending on -

DR. ECHOLS: Okay. So, there are really two possible answers?

DR. GILBERT: Yes, sure, if you've got a blockbuster drug and think you can do it with an active control, that's great.

Tom or Ed, did you want to amplify my remarks?

DR. FLEMING: No, I agree with your statement. The essence here is to look at what would be the most reliable way to go forward, to understand benefit to risk, to understand adequate evidence of safety, adequate evidence of efficacy, which obviously

involves, what would be the right endpoint, what would be the nature of the design, superiority could be an approach, a non-inferiority could be approached, but how would you justify the non-inferiority margin?

The issue here isn't so much what the answer is. The issue is, what's the scientific reasoning? What's the justification for what the answer would be?

DR. GILBERT: And before Ed speaks, I'm hoping that our colleagues from the agency will, as much as the law allows, also speak to this subject.

DR. COX: You know, I think we've touched on the major issues. It is -- a lot of it hinges around what we know about treatment effect, what the appropriate design is, what the endpoints would be. To the extent that we can try and flush some of that out, that would be helpful.

DR. GILBERT: All right, we'll just start at this end. Rick, not a clinician, but

any comments that you might have?

DR. NOLTE: I'm really underpowered to comment on the major points of this. But basically, I think one thing that's emerged, that I do feel comfortable speaking about, and it's in the area of diagnostics and improving our ability to identify those patients in clinical trials that could benefit from the drug, and the tools are there.

I mean, we've -- several speakers have touched on new approaches to diagnostics, better application of existing diagnostics, looking at other specimen types, other than sputum, those sorts of things. I think that's key in all of this.

The problem becomes, when you start talking about the newer technologies, the concept of companion diagnostics that we brought up. There's really -- although there are specific reagents and research-use-only reagents that are available to accomplish some of the things that we talked about today,

there is not the clinical efficacy data on the diagnostic side that supports their use.

So, you really have to -- that is going to have to come through some sort of partnership between the pharmaceutical industry and the diagnostic industry.

DR. GILBERT: Very good and I also meant to say, we hope you'll stick to a two to three minute limit here on this part of it. Ed, you certainly did. Thank you very much. Tim?

DR. MURPHY: So, I have three things to say. The first thing is that antibiotics work for community acquired pneumonia. There is the Austrian and Gold data, that I think shows a dramatic effect for penicillin.

We know that pneumonia is caused by bacteria in the lung. We have anti-microbial agents that are very active in-vitro. We have anti-microbial agents that are very active in animals and I take care of patients, they come in, they are coughing, they have infiltrates

on their chest x-ray, they have fevers. You give them antibiotics and they get better.

So, the question is not really, do antibiotics work or not. I think the key question is, how are we going to assess new agents, which we need in treating community acquired pneumonia?

The second point I would make is what I made at the end of my talk, is that I don't think placebo controlled trials are appropriate for community acquired pneumonia because I think the majority of community acquired pneumonia is caused by the pneumococcus.

We have effective therapy for pneumococcus and perhaps, most importantly, pragmatically, it's not going to be possible to enroll people in placebo controlled trials because IRB's are not going to allow and physicians are not going to want to do it.

It might be possible to take the very mildest community acquired pneumonia as a

select group and do it, but I think then we get into all the problems of dividing things up and not having meaningful results in that point.

So, the third thing I'd like is, so what do we need to do to address these issues, and I think for the long term, two key things, I my mind, would be better diagnostics.

Community acquired pneumonia is not one disease. It's multiple diseases. Pneumcoccal pneumonia is different from Mycoplasma pneumonia, clearly. So, if we had better diagnostics, we could actually design better trials and get better answers.

The second is, I think we need validation of patient reported outcomes. I think that will allow us to better trials particularly for community -- mild community acquired pneumonia.

The immediate thing, what should we do, I think it's critical for us to figure out a way to come to a consensus to design a well

done, non-inferiority trials or well done superiority trials with patient reported outcomes. As a whole, that's why we're here, we need new antibiotics and we need to come to a consensus using the best science that we have, whether it's Roger's way or the Cubicin data. Get the best numbers we can, decide on a margin and proceed from there.

DR. GILBERT: Thank you. Tom?

DR. FILE: First of all, let me just thank everybody again, for allowing me to participate in this. I've really learned a lot and based on what Tom Fleming said this morning, I hope my comments are not meaningfully worse than those of others.

But at any rate, I'm going to use the scenario of the patient you presented earlier, and to answer the questions, in that patient, the 35 year old who clearly has air space disease on chest x-ray, who has fever, who has leukocytosis, who has underlying co-morbid condition, who has a family, who has

probably got some illness, I mean, he's got all kinds of potential epidemiologic clues for either pneumococcus, haemophilus, mycoplasma, psittaci because of the parakeet, so he's got all kinds of potential clues there.

But I think that patient, it would be inappropriate to not treat that patient and use a placebo controlled trial. I think that patient clearly will benefit from anti-microbial therapy.

I think that it's unlikely that we can -- in superiority trials, if we use effective controls and using the standard types of outcome measurements that we've used in the past, we're ever going to see any difference, if we use the good effective controls, and I think what we need to do is evaluate some of these other outcome measures, whether they're biologic markers, such as pro-calcitonin, whether the patient -- response outcomes and looking at the speed of recovery or the time to event -- resolution of event, I

think that's what we need to do. Then looking at the pharmacodynamics to help predict as well.

In fact, I might use this, just to make a correction in my presentation. I think when I was reporting the response rates of that levofloxacin versus ceftriaxone/cefuroxime plus or minus erythromycin study and trying to just look at the patients who received oral therapy, thinking that that's sort of a surrogate for mild pneumonia versus pneumonia requiring intravenous therapy and acknowledging that about half the patients enrolled in that trial only received oral therapy, I said that the difference was -- I think 95 percent versus 88 percent.

Actually, it was 96.4 percent for levofloxacin in the orally treated group versus 89.7 percent for cefuroxime, which was statistically significant, but it does bring up this point that it's almost like the

daptomycin issue.

Subsequent studies, for example, from Victor Yu and Keith Klugman's group, when the looked at least at bacteremic pneumococcal pneumonia, showed that cefuroxime was pharmacodynamically -- well, it was clinically worse, and then they correlated it with the pharmacodynamics, showing that the drug, compared to at least ceftriaxone, does not have a good pharmacodynamic profile.

So, maybe there is another reason that helps explain that result that I just mentioned and I didn't have time, also, to present the 750 levo data versus 500 levo data, which showed that in the 750 arm, again, looking at the difference in pharmacodynamics, that the 750 actually showed quicker resolution of symptoms, at least if you look at fever, and then the third study, again looking at pharmacodynamics, that I presented from Jerry Schentag's group as well, that when they looked at the AUIC of the -- the

pharmacodynamics -- the macrolide that was used for the pneumococcus, they could predict what patients that failed as well.

And so, my final comment is, I think we can strive, maybe, for superiority trials, but in these studies, accept a -- maybe not reaching superiority, but if we reach a non-inferiority, lower limits -- bound of the 95 percent confidence interval, that's very acceptable, within 10 percent or whatever, that that would, to me, still be very acceptable.

DR. GILBERT: You're a great warm-up act for Dr. Ambrose, who tomorrow, will present the PK/PD data. Thank you, Tom. Robert?

DR. O'NEILL: Yes, I've been trying to integrate all of this great presentations that we've had, in terms of what can be helpful in terms of design, and the way I'm thinking about this is, it's been said that CAP is a continuum, and it's probably true.

But the struggle we're having is the problems that are especially therefore mild, because I think that mis-diagnosis is really more of a problem in the mild, than it is in the severe, and that comes with a number of issues, in terms of impacting the treatment effect.

Probably the treatment effect can be argued to be a smaller or more modest in that group, than it might be in a severe.

But I think that the solution to any new improvements in the design, whether it's a non-inferiority or a show of difference trial, is in better endpoints that take advantage of modern diagnostics, so that you have a more sensitive and specific outcome and it's -- that is responsive to therapy, coupled with better entrance criteria, which is also taking advantage of the diagnostics, in where you're essentially eliminating those folks who are going to get no benefit, but all the risk.

Then finally, I think thinking

about endpoints, that might start to take time to event and time to benefit into account earlier, rather than taking 21 day benefit, which essentially, if you've mis-diagnosed it, everybody is going to have a 90 percent improvement rate.

So, the whole Karen Higgins summary is the fact that if you are taking -- if you got mis-classification in a mixture population of folks who actually don't -- everybody is going to be better at 90 percent. So, that's a problem you're dealing with right now.

So, you're left with a situation of, are these equally effective or equally ineffective, because everybody was going to get better at 21 days. So, you've got to use some kind of endpoint that discriminates and it doesn't have to be a superiority trial against placebo, it has to be a discrimination trial against some other control or some other conditions of use.

DR. GILBERT: Thank you. Bruce?

DR. PSATY: I want to make four points. First, I think that overall, to improve these trial designs, the improved diagnostics will go a long way to helping the trial design, so that there is a homogeneous patient group, with a specific condition that can be addressed, and then that actually should be implemented in clinical practice and used.

As a cardiovascular epidemiologist, I prefer ITT in superiority trials and I confess that bias. I'd like to see efficacy and safety treated comparably, and they are in ITT analysis, so that you can get a good risk benefit assessment there.

It's important to use the randomized trial as a way to identify adverse effects and not to rely on investigator associated decisions about whether it was related to the drug.

Insofar as it's possible, pre-specified safety endpoints, as they arise from

safety signals, need to be identified and included in clinical trials, even if they're phase four trials, and when those phase four trials are done, they need to be done in a high quality way, so that we actually have, as a practicing general internist, the information we need to make use of these drugs.

In terms of the outcome, I'm okay with an outcome that involves clinical judgment. That's what we do for MI trials. In a sense, we have adjudication committees that decide whether the endpoint has occurred or not.

It's key that it's blinded, so that's a key methodologic issue. I'd like to see the patient outcomes incorporated. I think it's difficult because it's still, in my view, something of a research activity and we don't know what some of those patient outcomes mean, quite yet, and how they relate to the other outcomes we've used.

So, part of the difficulty we have here is that there's some basic clinical epidemiologic research that needs to take place.

Time to resolution could be a -- I have one of these symptom questionnaires, could be an attractive outcome. I would like to see placebo controlled trials. I'm not sure they belong in the FDA and in the regulatory environment. It might that the NIH needs to carve out a section of the community acquired pneumonia trials and see if there's actually a benefit there and that's an important activity. If it is, then we don't talk about placebo controlled trials anymore. That becomes the standard of therapy.

So, if there is an area, I'm not sure it belongs here, between the regulators and industry, to make -- to come up with that estimate at this point in time. I think we're -- the antibiotics are out of the bag. It's like devices, they're out there.

And so, I'm not sure we're going to -- that this is the place to insist on placebo controlled trials, though as Tom Fleming has pointed out, we really need that information to make intelligent decisions about what the non-inferiority margin might reasonably be. Thank you.

DR. GILBERT: Dr. Temple?

DR. TEMPLE: There's a lot going on. It seems -- many of the things that have been talked about, such as better diagnostics and use of patient reported outcomes and things, don't really help you in the non-inferiority setting because you don't have any better data on what the effect on those things is, than you have anything else.

So, they could help you do a superiority study, but I don't hear much in the way of superiority studies being proposed, unless we can do what Bruce wants to see, and actually use placebos. All the scuttle-butt I hear is that people are not going to be

willing to do that. They're barely willing to do it where there's very little evidence of benefit, such as sinusitis. I can't imagine they're going to do it here.

So, the most important thing, it seems to me, if we feel we need studies in mild to moderate disease, and I'll come back to that, is to see if we really can -- for example, by looking at the three day period or something like that, identify an effective treatment that is clearly larger than what a no-treatment group would get and then we can do the trials and there's no problem.

As people said -- I guess, Bob said, you wait until 21 days, you probably have very substantial improvement, even if you didn't have any effect.

But in all of these things, we have to be able to say what the drug did. The fact that people would be satisfied with a difference of five percent is of no consequence at all. They are perfectly right,

they would be satisfied with a difference of five percent and that has nothing to do with whether the drug works.

Our problem is to find out whether the drug works. We don't care if the difference is too small to be of interest to practitioners. We've got to know it works, otherwise we can't approve it.

A question that I think ought to be considered is whether you actually need information on all severities of the disease, if you had solid data. If we knew for sure that in bad disease, we could define the effect size of treatment and we had rock solid data on very severe disease, do we actually have to have information on all stages of the disease?

For what it's worth, in hypertension, we label drugs as lowering the blood pressure. We don't particularly worry about how severe it is. That's the clinical decision people make based on JNC-7 or

whatever.

So, that's worth thinking about. But if you could do it, it would be certainly good and you'd like to know that you really were having an effect.

Just one matter on the way we look at safety. It's true, there's usually no prior hypothesis in safety. The way that it's dealt with, however, is that we believe everything. So, we don't cross things out because multiplicity -- if you corrected the side effect data in trials for multiplicity, you'd never have a significant finding, but we don't do that. We put it all on the label anyway, as if it's probably true, knowing perfectly well that some of the things we find probably are not true and are the result of multiple observations.

When you have a hypothesis later, then as people said, you want to design a trial that really can do some good.

So, the thing I heard here that

most interested me is the possibility that if you look early, like at three or four days, you might in fact, have very good evidence that treatment is beneficial and if we all come to be satisfied with that, I think we have an easy resolution of this problem.

But not everybody agrees that that's a lock, yes. I don't think John Powers is convinced yet. But we need to look closely at that.

DR. COX: I'll just make a few comments. I think one of things we heard earlier in some of the question and answer period, was the issue of prognosis versus benefit, and I think this issue sort of intermingled with that of enrichment and are there things that could be done in the mild to moderate community acquired pneumonia population to further get a population where there might be more benefit.

I recognize in part that that's conceptual because the question is then, how

do you understand what the benefit is in that group, which is one of the challenges I think we're all struggling with.

With regards to superiority trials, I think we've heard a lot from a number of folks there and it does sound like it's an area where it would probably be fairly difficult for most anti-microbials to show superiority, unless there is development of newer endpoints or different timings of assessment, that may help to discriminate one drug from the other. But it does seem that it's -- the demonstration of superiority would be a real challenge.

Then beyond that, the comments on -- we heard some about looking earlier -- this gets to the issue of timing of assessment, and certainly, there's more to be done to look there, to see if there is the possibility that that might be an endpoint that may help us to further understand treatment effect.

Then, again, a point that's been

made from the discussion so far, we've heard -- and understandably so, because it's difficult to understand which patient, even with mild to moderate disease, might progress. The issue of a trial where patients were to get a placebo -- from the comments we've heard today, I think a number of folks have expressed some degree of concern over that because of the possibility that some of those folks might progress and that's an understandable consideration, and those are the comments I have. Thanks.

DR. FLEMING: Thank you. Just to begin, thanks to all for what has been an extremely informative day, lots of issues out. What we wanted to do was to try to get all sides of the arguments out, and at least, we made an attempt that got us, at least, part way there.

My own sense here is what's -- where to start here is, what is the endpoint? What's the active comparator? What -- that's

going to have so much influence on the design, and we've heard the discussion about this from many people, Dave Gilbert, Karen Higgins, Tom File, are amongst those who have given a lot of specific insights about that.

What I've heard quite uniformly from the collective presentations are that a clinical measure, a measure that unequivocally reflects tangible benefit is one that's quite strongly supported.

Now, there's not a single proposal for what that would be, but the kinds of components or aspects of that, that I'm hearing are resolution of key symptoms, symptoms such as cough, shortness of breath, chest pain, returning to work, usual activities, and of course, issues like hospitalization mortality, but those are unlikely to occur in a mild setting.

Time to those events certainly provides a great enhanced sensitivity. If we're looking at a scenario where 90-odd

percent of people will have resolution, seven to 21 days after end of therapy, then certainly a time to event is going to give you an enhanced sensitivity.

There's a lot of exciting discussion about PRO's and I strongly am intrigued by that and supportive of that.

A key point here is what does it have to be then? What's the control arm, to show an effective intervention on one of those measures? Without question, the most reliable interpretable data would come from a study that would show superiority.

Can we do a non-inferiority trial? I guess one point that needs to be made up front is, there's no such thing as a non-inferiority margin that would apply to all endpoints and all comparator arms. Each separate combination of comparator arm and endpoint needs to have a separate justification for a non-inferiority margin.

I've heard only one, it came from

Roger. It came from his interesting discussion today about one possible way of justifying a non-inferiority margin that would be data based.

It's an interesting argument, but one that I'd say is pretty fragile and does need to be explored. It is certainly worthy of exploration, does need to be explored in much greater depth. It seems to be based on the argument that if you've got 35 percent of your population here, that are the CAP with atypical pathogens, that these are people that would do very badly without anti-microbial and will do extremely well with anti-microbial. Is 35 percent the right fraction in our trials?

The argument that he was giving is the non-successes aren't necessarily all in that group, but if we put them all in that group and thirdly, if we allowed for the possibility that more than 10 percent of those people would, in fact, have a response in the

modern day age on a placebo, these are all issues that haven't been considered in that argument and would greatly erode what you would come up with, with the non-inferiority margin.

But just on this last point, how strong is the evidence that those patients that are CAP with typical pathogens would, in fact, be people in today's era and today's interventions and today's assessments, that would, in fact, be failures according to our endpoint? It's based on the Bullowa data from 1937.

The patient selection issues are different. Patients are different then from what we're looking at now. The supportive care is clearly different then from what it is now. The definition of the endpoint is certainly not necessarily consistent then from now.

We're looking at test of cure, seven to 21 days post-treatment. Well, what

was 21 days post-treatment in the Bullowa data, in terms of that long term outcome? Is it really going to be 10 percent or might it truly be something much more substantial?

So, when all of this is coming together, this is an issue that I think is worthy of further exploration, but there are an awful lot of issues that are fragile here, around what is, to my way of thinking, the only data that's been put forward to justify an non-inferiority margin on some endpoint, an endpoint that in fact, might not even be the one that many of us would in fact, view to be the most preferred endpoint.

So, bottom line is, the clinical endpoints that are being suggested here are intriguing. None of them, with the exception of one possible exception, has had anything put forward that would justify what a non-inferiority margin would be.

Clearly, we will have a much clearer sense of benefit with the superiority

trial. Going non-inferiority is, as is always the case, a treacherous way to try to understand whether you're getting true, favorable clinical -- true favorable benefit to risk.

DR. GILBERT: The Infectious Disease Society got into this because of our concerns, obviously, for approval of safe and efficacious drugs, and also, by the decreasing number of drugs that are in the pipeline.

Discussing this with colleagues at the FDA and the industry, the feedback was there was uncertainty and we've heard a whole day, trying to address that uncertainty, and I'm hoping that we're getting closer to reducing the levels of uncertainty, so as to continue or spark the interest of industry in developing new drugs.

So, one thing I've heard in the way of uncertainty today is, placebo controlled versus no placebo controlled, and as -- if I put on my clinician hat, I can certainly

understand that. I can understand the anxiety about not having a drug on board that was active against the pneumococcus.

I cannot see any trial being approved, at least in our Institutional Review Board, unless you could at least cover that bonafide pathogen, and we heard from Tim, that we're probably underestimating how much mild community acquired pneumonia has pneumococcus included.

So, I think we need to get innovative, in terms of clinical trial design. One way might be a placebo controlled trial with a rescue arm, if the patient is failing after two days or three days, whatever the appropriate time interval is, and then you can implement a drug that has activity against the pneumococcus.

I'm pretty comfortable in not having an active drug against mycoplasma and chlamydia, pneumonia. They cause morbidity, but they're not life threatening infections.

Another possibility, which I mentioned to Tom -- I have to point out that Tom and I have had some very spirited discussions in the weeks leading up to this event, is that nobody has suggested a three arm trial, and I'm sure industry will now shudder and throw something at me, but if you had one arm, which was penicillin or ampicillin versus placebo, another arm that was penicillin plus macrolide, and if you had a third arm, which was your new drug, you would sort of cover all the bases. We'd learn a hell of a lot.

I mean, you'd have activity against the pneumococcus and no activity against the atypicals. In the second arm, you'd have activity against the atypical and the pneumococcus, even if the pneumococcus was resistant to the macrolide, and then you'd have whatever the study drug was.

I'm not going to reiterate what everybody said about patient reported

observations, etcetera etcetera.

Safety comments, one of my dreams is that one of the drug companies or multiple drug companies would truly cooperate in a very prospective way with the diagnostic companies and I mean, that seems to me, to be a win/win situation for everybody, but in particular, the patient population.

Lastly, I think we've got to get more active looking for post-marketing adverse effects. Finally, I'm -- I don't want to get too emotional here, but I'm at -- I feel abhorred by the fact that we're still doing passive monitoring.

We have incredible electronic connectivity with the world. The pharmaceutical industry knows that every time I prescribe a drug, every time Dan Musher orders something they know it.

Why can't we have a sampling of users on -- a very focused sampling saying, "Have you observed any unusual adverse

effects," on a prospective basis, instead of waiting until some huge outbreak occurs of livers that don't work and so on and so on.

Then the last thing is resistance. The only thing George and the others didn't mention -- I guess they sort of mentioned it in passing. We ought to be looking for development of resistance during clinical trials, rather than after clinical trials, routinely, absolutely.

DR. SPELLBERG: Well, I also got interested in this and became involved through the AATF and my boss, Jack Edwards, has a thing about emphasizing that the antibiotic problem is no one's fault. When I say the problem, the fact that we're getting less and less of them. It's not anybody's fault.

The relative parties involved in this process are all looking after, appropriately, what they're suppose to be doing and unfortunately, the result is this societal conundrum, where we need to maintain

safety and efficacy standards, but we do need to have new drugs. We desperately need to have new drugs.

So, I'll go back to a point that Dr. Bradley made earlier today, and I think that the statistical discussion has been phenomenal, and I'm somewhat awed by the brain-power on this side of the table over here, with respect to that, and these issues are critically important, of course, but I think we do need to find a balance between what's practical and achievable and what the statistical evidence will support in a trial.

I am encouraged that some of the ideas that Roger brought up might be promising and agree, they should be vetted more thoroughly, but maybe that's the direction we need to go.

DR. LAESSIG: Sure, I'll take a brief moment to comment. For those of you who don't know me, I'm Katie Laessig, the Deputy Director in the Division of Anti-Infective and

Ophthalmology Products FDA, and it's been a fascinating day.

I have a few things to say, they are my opinions only, and I'm going to limit my comments to mild pneumonia.

At this point, I don't feel that we've adequately described the treatment effect in mild pneumonia. Therefore, either we have to somehow extrapolate from moderate to severe, which is the topic of tomorrow's discussion, or find the will to conduct placebo controlled trials.

I don't agree that they are necessarily unethical. I think a carefully conducted trial and carefully selected patients, with scrupulous monitoring -- and as Dr. Gilbert mentioned, perhaps an early escape might be possible.

I also feel that, you know, the assertion that it's unethical just because we believe that there is a treatment effect, even though we don't necessarily know what it is,

does not hold a lot of water with me because you may be prescribing antibiotics for something for which patients are not benefiting and it is contributing to anti-bacterial resistance, and that's it.

DR. ECHOLS: Yes, thank you. I'm going to jump around just a little bit, because I don't want to be redundant. I just wanted to start by saying, industry, as much as it may be apparent otherwise, is not resistant to new ideas, is not resistant to new clinical trial designs.

The issue has to do with, what are you going to get at the end of the day? What's the risk of these new study designs? To jump from one pan to another, without knowing what's in between, I think, is still why industry is very conservative.

What I mean by that is, it's not so much that we don't want to have PRO's. We've incorporated PRO's into our sinusitis study, into our AECB trial, we just don't know how

they're going to come out, and this was illustrated, I think, beautifully and these data were presented, we took the PRO for acute otitis media, that was developed by the group in Pittsburgh, Hoberman, et. al., and we implemented that PRO in our phase two trial.

However, our phase two trial introduced something else, and that was tympanocentesis, and the fact that you were sticking a needle in the ear of the kid at baseline, all of the PRO scores went `sssss', in the first eight hours.

So, it totally obliterated the value of the PRO because we introduced something that was required from a diagnostic point of view.

So, there is just a lot of variables that we don't understand. So, I'm all for introducing PRO's, but you can't make the jump a priori that this will be a good way of either demonstrating superiority or demonstrating -- even what the correlation is

with other outcome measures.

I will never be involved in a placebo controlled trial in CAP. And that's my experience and unless we can eliminate the possibility of pneumococci, I wouldn't do it, and if I eliminated the possibility of pneumococci, I wouldn't want to do it. So it makes no sense to me to try to show superiority over placebo. It's not just an ethical issue to do the study and design a highly selective group of patients, are not the type of patients I want to get labeling for. Makes no sense.

Again, I think the issues of validating PRO's, there are two steps in validating PRO's. One is all the construct validity and are the questions reproducible, and do people understand them and all the rest of that. But then you have to see how sensitive of a measure are they in clinical trials. And we don't have that information and so to try to design and say it's a more

sensitive way of showing a difference between two active treatments. I think again, we can't assume that that's going to happen. We should be investing in our clinical trials, we should be incorporating new outcome measures. But we can't take those until we know what they really show.

Finally, I think we really can make our clinical trials better. I've been involved in these things as an investigator, since 1979 and on the industry side, since 1989.

An awful lot of our studies, I would say, are just awful, in terms of patient inclusion, what are we really looking at? I have no qualms. I'm not debating the fact that there's a lot of background noise and when you apply that in a non-inferiority setting, you don't know what you have at the end of the day.

But what I do think is that there is a treatment effect in typical bacterial

pneumonia, and I think what we should be doing -- our effort should be -- at least, if you want to develop a drug for the treatment of pneumococci or Haemophilus or Moraxella, we should be focusing on how you select patients that truly have bacterial pneumonia, and you can do that with improved clinical inclusion criteria. You know, require fever, require sputum.

Now, you're going to eliminate some patients, because not everybody has fever, not everybody is able to produce sputum, but you're more likely to get rid of some of the background noise if you're more restrictive in your inclusion criteria.

And then the better diagnostics, I think can help, but right now, they're not ready for inclusion/exclusion criteria. They might be good for post-hoc analyses, but they're not there for screening purposes with rapid turnaround, where patients are enrolled, particularly in mild to moderate pneumonia,

and heaven forbid, you not start treatment within four hours of that patient hitting the emergency room, because then you'll get a demerit on your scores as a clinician and your hospital gets demerits as well.

DR. GILBERT: Thank you, Roger. George?

DR. TALBOT: Yes, I have comments in three areas. The first relates to our definition of severity, the second relates to the treatment effect and what we know about it and the third relates to answering your questions. So, I thought I'd leave that until last.

With regard to severity, I want to reiterate some of the points that I tried to make, perhaps not in a very articulate fashion at the beginning.

Our discussion of design and enrollment criteria has been framed, I think, in two different ways and this follows on what John Powers was talking about, and two

different perspectives regarding the patient populations that are suitable for study, to answer this question.

The first perspective is the clinical care approach and the words we see around that are requiring hospitalization. Things that we've mentioned are somewhat vague and subjective.

The second approach, which I think is needed if we're going to answer some of the questions we've asked here today, is the approach of what do we need for clinical research? I think we still need to be more precise and more accurate in our definition of severity, within what we know. The best example I could give is why are we continuing to lump mild and moderate. I'm not even sure that one person could say mild -- that a patient might be mild or moderate, who knows? So I think we need better validated tools for defining severity before we do these studies.

One could say what about the PSI?

Well that's a wonderful tool, but I was thinking what John Retts mentioned as well, this was derived to predict whether therapy could be given as an inpatient or an outpatient. So they all receive therapy. So, as a result, I don't think we really know what mortality would be or what morbidity would be in any of the PORT classes in 2008, for sure. We can guess that one, two, and three might be pretty low in terms of mortality, but I don't know that the inflection point, untreated, in mortality, is the same as the inflection point treated, so in treated, the inflection point is after three. In untreated, it might be after two. And you also might progress from one to two to three. So I have a great deal of concern about using the PSI score alone. So I think a potential solution there is to use, as we discussed today, the PSI plus some other characteristics and/or diagnostic tests that would give us a much more precise estimate of severity of

baselines, so we know who's being enrolled. So maybe PSI plus frail plus PCT. So I think that that's essential if we're going to answer the question in a scientific way, especially if we're going to go into a potential superiority study.

Now, in terms of the treatment effect, I don't think I'm going to be sitting here tomorrow, so I'm going to get my two cents in. For severe typical pneumonia, I think that there is a treatment effect that's appreciable and supports a Delta of 10 percent, and that's based on data from the pre-antibiotic era, the animal data we've referred to, the high mortality that's seen in "severe patients." I think that an NI approach in carefully conducted studies, is appropriate for those patients.

For mild disease, whatever we define that is, I would say that yes, there probably is a treatment effect, but it's really hard to tell what it is and I would say

that it's probably reflected primarily in time to resolution and so, if we're going to study it, we would want to see, okay, does it take three weeks to get better without antibiotics or versus two days with.

So, there, it's possible we could define an NI margin. It's possible that might work, but I'm uncertain on that, and then that's my intellectual honest truth. But we need an answer fast.

Finally, what about designs for mild CAP? Well, first of all, we have to be sure that it's mild, to begin with, and so, I'm willing -- since I'm not sure about an NI margin, to entertain the possibility of superiority trials, and that takes us down, do we use placebo or do we use active?

With our current state of knowledge, I have real reservations about using placebo-based superiority trials, even in mild, I think we need more information there and maybe take it step-wise. If you

took a set in the mildest of CAP, the mildest of the mild, and as Dave mentioned, well there may be an early escape, in a hospital setting, might do that. But I don't know that that's Pharma that should do that. I think that might be NIH, for example.

So, I'm not ready to go to placebo in mild, except maybe in that very specific setting and maybe not by those around the table.

An alternate approach could be to do an active study in mild and maybe there, you do two things. You do active for three days versus five. You do active for two days versus five or seven. Some of those things have already been done, but that might give you a window on the importance of time, as we've discussed.

It might also be possible to not use an escape approach, but to do active at base line in one, and then active after 24 hours in the control arm, or something like

that, with the patient hospitalized for that first 24 hours.

So, I think we do, as Tom mentioned, need to think creatively about ways to answer that, but the people who do that study may not be in this room.

DR. GILBERT: Thank you, George.

DR. BRADLEY: Most of my comments are going to relate to pediatrics, but certainly, there's a lot overlap with adult considerations.

In society and by the FDA, children are considered a vulnerable population. So, how we view them is a bit different for clinical trials than adults are viewed.

Like others have said, I think entry criteria into these studies really need to be tighter for pathogens, and it's nice to know that there are new diagnostic techniques. In children, it's reported that 90 percent of community associated pneumonia is viral, so it's even more critical with children to know

if you're really treating a bacterial pathogen or not, and we generally don't do lung taps in children. So, those sorts of invasive procedures are frowned on by IRB's. So, there has to be other ways that we can be creative to get a diagnosis.

In addition, you want to make sure that only those children who have the real disease get exposed to investigational drugs, to limit unnecessary toxicity to children. You don't want to be exposing children with a viral pneumonia to a potentially harmful antibiotic.

I believe somehow, that non-inferiority trial designs need to be the basis for drug approvals and I know that that's a difficult concept and we've talked around that a lot. I'm reluctant to use a placebo in a drug, in a trial that's looking for approval of a drug.

Certainly, there are places in the world, say, for otitis media, where

antibiotics are not used initially and there may be places that the NIH can perform studies, where the ethics of not using an antibiotic for pro-calcitonin less than .25 might actually be effective and the NIH has funded studies in Scandinavia that we use for drug approvals for pediatrics. So, that's a possibility, where you can get natural history information on mild, moderate and severe CAP, or at least mild, where it's ethically feasible.

The NIH would be a place, if you wanted to study that in the United States, that I think would be the more appropriate funding source and they, indeed, right now, are funding a study in children over two years old with otitis media in a placebo controlled trial in Pittsburgh.

So, the concept that we can get information from the NIH, as opposed to from the pharmaceutical industry, I think, is a very relevant one.

I believe also, to focus on -- if we're doing non-inferiority, to focus on what the appropriate Delta is and there's been a lot of discussion on that, and to define what a meaningful benefit is and we've been through this argument in pediatrics, what an epidemiologist believes is a meaningful benefit will be different than the regulators, will be different than physicians taking care of patients, will be different than the parents of the children.

Is a half a day or a day improvement, in a natural history of a disease that's four or five days, enough for you? And the parents will all say yes, and epidemiologists will say no, so there's got to be some consensus of what society wants.

Also, we tended to use the adult Delta for pneumonia for pediatrics and I don't know that that's the right thing, because -- kids are smaller, we need a smaller Delta?

(OTR comments)

DR. BRADLEY: The five year old will probably have a -- more -- a faster spontaneous resolution of pneumococcal disease than a 30 or a 50 year old. I don't know this. It's biologically plausible. It needs to be tested. A six month old may take longer than a five year old. A two year old, where does the two year old fit in?

And to ask industry to do 2,000 patient trials at a six month old, and a two year old, and a five year old just are not feasible. Maybe the NIH can fund that study.

There is also an interesting source of information to look at what an appropriate Delta is and what the natural history of untreated disease is, that may be in the FDA's database.

While I was on the Anti-Infective Drug Advisory Committee, we're certainly aware that antibiotic approvals come to the agency and all that information is highly confidential, kept at the agency and not

public, until it's presented and not all the data is presented.

If the drug doesn't work, the company doesn't want to come in front of the public and say, "Our drug didn't work." Yet those data are likely to be at the FDA. So if there's a company that picked a wrong dose, so that there's no effect, that could be closer to a placebo effect. And in that trial we may actually have some information hidden that gives us more insight into what a placebo will do. And I know there's going to be issues of confidentiality and that the data will have to be put together in such a way that that drug's not named, but there's got to be an incredible amount of data within the agency that's not in the public domain that can actually help us figure out a Delta so that we can improve on mortality, as well as morbidity.

Finally, one other piece of information, we were part of a levofloxacin CAP protocol and of course, the fine criteria

are for adults, so we put together a fine criteria -- a modified fine criteria for children, which of course, is not validated. But for children under five years of age, we had as a comparator, amox/clav and -- which has no ostensible activity against Mycoplasma, and in this era of pneumococcal vaccine for children, we had very few pneumococcal pneumonias. Most of our pneumonias, even in kids under five years of age, were Mycoplasma. And if you look at the efficacy of amox/clav and levofloxacin, they were the same. And this was just published a few months ago. But this is in kids under five.

Now, is this relevant to the 18 year old, the 30 year old, the 50 year old, I don't know, but it's very interesting information. Thanks.

DR. GILBERT: Thank you. John, last thoughts.

DR. POWERS: I'd like address some of these issues about putting together the

evidence for looking at whether a non-inferiority trial makes sense in this setting. But I'd like to start off with two general points.

It often seems stated that we come back to this issue of trying to split out statistical and clinical issues, and in fact, appropriate clinical trial design is not merely a statistical issue. Statistics is a way of evaluating the precision of what you're actually looking at, and when I look through this pile of information from the old studies, there is very little statistics in this, actually. It's mostly case descriptions of what happened to people in the past.

The second issue is, we've talked about risk to patients and obviously, it's important to talk about risk of not giving a drug, but there's also the risk of giving a drug, when we're unclear of the effectiveness of that drug and we may all choose to believe something, but science isn't based upon

beliefs. It's based upon what that actual evidence is.

The third thing is that one of the other scientific tenets, and I think, Tom, you showed it in one of your slides. You had a single line for the effective placebo, but underneath it, you wrote `meta-analysis of effects', and that means looking at the totality of the evidence that's out there, not picking perhaps, one of the studies that we like the best and evaluating that one.

So, I can pick up this one from Davies in 1935, where they say it's important to keep in mind, the probable benign course in a large proportion of the cases. The not infrequent early crisis, and the youth of the average patient, we based our impressions upon relief of symptoms, the fall of temperature, the day of crisis and the speed of resolution and the incidence of complications.

So, try to find out what the percentages are in this paper? You can't.

It's impossible. So, looking at that, then what we're really doing in the absence of placebo controlled trials is, we're trying to evaluate what you can call historically controlled trials and relate them, in a historical way, to another historical controlled assessment, which is the non-inferiority trial today.

So, what does E-10 say about where can you use historically controlled data in the most rational place? It's one, when you have objective endpoints and you're very clear on what those effects are in a reliable and reproducible way.

So let's take Bullowa, who looks at sustained or substantial improvement, and how do we relate that to what we're measuring in today's trials? I would argue that we don't know what either one of them means. We don't know what Bullowa was using as an outcome assessment, neither are we clear today on what a clinician's judgment that the person doesn't

need any more drug means today.

So we also know there are several studies that actually look at, historical controls on average underestimate the effect of the control in the historical evidence. So we're widening that by the mere fact that it's historical, and Steve Snapinn has a really good article on what are some of the issues with our non-inferiority trials. All the data we saw that Karen Higgins presented, all comes from non-inferiority trials, where people may overestimate the effects of drugs because they know, even in a blinded trial, that everyone's getting an active intervention. Which may make them code, even borderline cases, say "Oh I know he's getting an active something, so he must be ok."

The other issue is that in our current trials, there is no requirement that someone be designated as a failure at day three at all. If you are designated as a failure at day three, you're carried forward.

But someone could get better on day five, six, or seven because most of our trials have 10 to 14 days of treatment and still be coded as a success. Therefore, if we compare a point estimate of effect today to a point estimate of placebo in the past, we're comparing a longer duration for today's current trials to a three day assessment of placebo in the past. All of which makes this all very problematic to be able -- Do I wish there was evidence for this? Absolutely. I didn't spend time reading this pile trying to not find anything, we're actually trying to look for it. And I agree with Bob, I actually think that the most rational thing here is why don't you study severe disease where we know there's an effect, and then actually show that your drug is effective in that setting.

Roger brought up a really good point. We have this idea that everybody needs to get intravenous therapy for severe disease. But in fact, we have very little evidence to

support that assertion at all. And most of these drugs had great bio-availability where the oral drugs could be studied in that setting as well.

The other issue I want to talk to you on, on placebo controlled trials and superiority trials. Janet Wittes came to NIH a couple of weeks ago and she gave a talk on the womens' health initiative study that compared estrogen and progesterone to placebo. She told me something I didn't know. And that is that the August Institute of Medicine called that trial unethical, overly expensive, and inefficient before it started, and they had to do a lot of wrangling to even get that trial off the ground. And that trial had something like 27,000 people in it. It was enormous, that trial. But Janet said something that was very interesting. She said, "You know when they asked me to be on the Data Monitoring Committee for this, I thought I already knew the answer too. The

only problem was, I knew the wrong answer." So the question is, if people believe from 30-some odd observational studies that hormone therapy had an effect in that setting. And what I get kind of discouraged about is some of the evidence I've showed you that in the past, it was ID trials that actually moved the whole science forward. And now I hear us insisting on belief instead of evidence and saying we can't do these.

So how can we do this? E-10 also talks about where in a more severe disease -- what could you do in that setting that's a superiority trial and it talks about dose response trials as an option in that particular area. Paul Ambrose is going to show you some data tomorrow. And Paul I don't want to steal your thunder on this, but also it gets to the issue of when Paul is going to show this that if you extrapolate to what people who had exposures of zero in their bloodstream would be in a current trial, it

comes out about 75 percent. That's a far cry from 1.3 percent, isn't it?

So that actually makes you think, well how can we compare yesterday to today. Now that's an extrapolation where you're just drawing the line, so that's got its own issues as well.

But the other issue before we leave placebo controlled trials behind as well, there's the issue of, what is the cost of failure? I like Dave's idea of, if you're going to do a placebo controlled trial with an early escape, what would happen to people? We heard all morning that the mortality in people with mild disease is next to nothing. Right? And that's not at a two day assessment either, that's further out. The Pneumonia Severity Index used 30 days as the mortality in that.

What would happen to people if you witheld therapy for two days? They'd cough for two days more? So, again, if we knew what

the effect was at two days, you wouldn't want to expose people to that harm. But in this setting, we wouldn't be exposing people to that much of a problem if we then just did that delayed therapy as well.

Last point is, I heard several times when I was at FDA, something that really bothered me about, well if it's a superiority trial, we don't have to worry about the other design issues of whether the person has the disease, etcetera, because it's a superiority trial, and the risk is on the sponsor if they don't do it very well. That's true, but the risk is also to the subjects who volunteer for the study, and if it's a superiority trial that tells us the wrong thing by a mis-classification bias or whatever, we then expose a lot of people.

Even bigger problem, if it does show an effect, how's the next non-inferiority trial going to be designed? Just like that trial. So we carry forward those errors and

flaws in those trials. So I think if we're going to do superiority trials, we have to make sure it's not just about a margin, we have to make sure those trials are done in a rigorous way. I'll stop there.

DR. GILBERT: Thank you, John. Well, we've been making suggestions about the NIH all day. Dennis, you're our NIH representative. Your thoughts.

DR. DIXON: I have very little to add to the comments that have already been made. I believe that the limitations of non-inferiority trials have been presented very thoroughly and in my view, convincingly.

One of the issues that comes up -- and the only thing that I'll offer a comment on at this point is that, in trying to specify a margin, if it is true that the standard is going to be a 95 percent complete response rate, what should -- then allowing a 10 percent absolute margin, that looks to me more like a 200 percent increase in the failure

rate, from five percent to 15 percent.

So, I wonder if we shouldn't really be talking about relative changes in discussing the margin and then, the 10 percent wouldn't look like such a modest margin.

I think that I've learned a lot today. It's been very stimulating. This is an area in which there are challenges that I haven't encountered in working primarily over the last several years in HIV disease, with the problems in identifying exactly what the study population is, because you don't know exactly what the pathogens are for all the volunteers at the time of entry, nor is it so obvious what the study population you would like to use, in order to do the most efficient clinical trials.

These are all -- there's always a trade off in here. No matter how you design the trials, there is eventually going to be some extrapolation required, to go from the clinical research to the practice of medicine,

because not every possible patient type will be represented in a clinical trial.

But there just needs to be a clear understanding, I think, of some of these trade-offs. How much do you want to build in a prospect of the need to extrapolate versus the limitations of trying to very narrowly define the patient population in any particular study.

DR. GILBERT: Thank you very much. I want to thank the panel and the audience. You've been patient. You've been involved. You haven't been hesitant to express your feelings, that's clear.

I can only promise you that tomorrow will be better. We'll start exactly at 8:00 a.m. because our goal is to get you out of here by 4:30 p.m. at the latest tomorrow. That means we have to start early.

We do have reservations downstairs, for those people that are staying in the hotel. I think we said we were going to be 25

or 30 people. That's a very flexible number. We hope you'll join us. Very informal, no formal agenda. See you tomorrow.

(Whereupon, the foregoing matter concluded at 5:42 p.m.)