DEPARTMENT OF HEALTH AND HUMAN SERVICES

FOOD AND DRUG ADMINISTRATION

CENTER FOR DRUG EVALUATION AND RESEARCH

CARDIOVASCULAR AND RENAL DRUGS ADVISORY COMMITTEE

NDA 21-359 Cellegesic
(nitroglycerin [NTG] ointment)
0.4% intra-anal, Cellegy Pharmaceuticals, Inc.,
for the proposed indication of relief of pain
associated with anal fissures

Tuesday, April 25, 2006

8:00 a.m.

Hilton Washington D.C.
North/Gaithersburg
Gaithersburg, Maryland

C O N T E N T S

PARTICIPANTS

Chair: William R. Hiatt, M.D.
Executive Secretary: LCDR Cathy Groupe, B.S.N.

Voting Committee Members
David L. DeMets, Ph.D.
Steven D. Findlay, Ph.D.
John M. Flack, M.D., MPH
Robert A. Harrington, M.D., F.A.C.C.
Frederick J. Kaskel, M.D., Ph.D.
Michael A. Lincoff, M.D.
Thomas G. Pickering, M.D., D.Phil
Ronald J. Portman, M.D.
John R. Teerlink, M.D.
Lynn Warner-Stevenson, M.D.

Temporary Voting Member (SGE Consultant)
Walter A. Koltun, M.D. (by telephone)

Non-Voting Participant
George S. Goldstein, M.D. (Industry Rep)

Non-Voting FDA Participants
Robert J. Temple, M.D.
Norman L. Stockbridge, M.D., Ph.D.

P R O C E E D I N G S

Call to Order and Introductions

DR. HIATT: Good morning. I would like to welcome everyone to the meeting this morning. My name is Bill Hiatt. I am the chair of the committee and from the University of Colorado.

I would like to start with introductions. Let's start on the lefthand side of the room here.

DR. GOLDSTEIN: George Goldstein, the Industry Liaison Representative for today, sitting in for Dr. Neylan.

DR. FLACK: John Flack, Chairman of Medicine, Wayne State University.

DR. DeMETS: Dave DeMets, University of Wisconsin, biostatistician.

DR. PICKERING: Tom Pickering, Columbia Medical College, New York.

DR. PORTMAN: Ron Portman, University of Texas in Houston, pediatric nephrologist.

DR. KASKEL: Rick Kaskel, Albert Einstein College of Medicine in the Bronx, pediatric nephrologist.

DR. FINDLAY: Steve Findlay with the Consumers Union. I am the Consumer Representative on this panel.

DR. WARNER-STEVENSON: Lynn Stevenson, cardiologist, Harvard Medical School, Brigham and Women's Hospital.

DR. HIATT: Again, Bill Hiatt, Vascular Medicine, University of Colorado.

LCDR GROUPE: Cathy Groupe, Executive Secretary for the committee.

DR. TEERLINK: John Teerlink, University of California at San Francisco and San Francisco VA Medical Center.

DR. HARRINGTON: Bob Harrington, Duke University, cardiologist.

DR. LINCOFF: Mike Lincoff, The Cleveland Clinic Foundation, an interventional cardiologist.

DR. STOCKBRIDGE: Norman Stockbridge. I am the Director of the Division of Cardio-Renal Drug Products, FDA.

DR. HIATT: Bob Temple will be here in a minute, and participating by telephone is Walter

Koltun. If you are on, can you say hello?

DR. KOLTUN: This is Walter Koltun. I am a colorectal surgeon at the Milton S. Hershey Medical Center, Penn State College of Medicine.

Conflict of Interest Statement

LCDR GROUPE: The following announcement addresses the issue of conflict of interest and is made a part of the record to preclude even the appearance of such at this meeting.

Based on the submitted agenda and all financial interests reported by the committee participants, it has been determined that all interests in firms regulated by the Center for Drug Evaluation and Research present no potential for an appearance of a conflict of interest.

With respect to FDA's invited Industry Representative, we would like to disclose that Dr. George Goldstein is participating in this meeting as a non-voting industry representative acting on behalf of regulated industry. Dr. Goldstein's role on this committee is to represent industry interests in general, and not any one particular

company. Dr. Goldstein is a retired employee of Sterling Drug.

In the event that the discussion involves any other products or firms not already on the agenda for which an FDA participant has a financial interest, the participants are aware of the need to exclude themselves from such involvement and their exclusion will be noted for the record.

With respect to all other participants, we ask in the interest of fairness that they address any current or previous financial involvement with any firm whose product they may wish to comment upon.

DR. HIATT: Thank you very much.

What is on the agenda today is an indication that is not typically discussed with this committee, but uses a drug that commonly is used in cardiovascular medicine.

Dr. Norman Stockbridge is going to give us a little background on this meeting. I would like to say, though, that I think in cardiovascular medicine, we do wrestle with symptomatic therapies

and how to assess them, so we will learn a lot from today's discussion.

Norman, could you give us a bit of a background?

Introduction and Background

DR. STOCKBRIDGE: I will just say a couple of words. I really do think that the reason why this is an appropriate venue to discuss this is that it really gets at some issues about symptomatic treatments and the drug was in our division, so it really did seem like it was appropriate to have you folks, who have thought a lot about symptomatic treatment in cardiovascular trials, deal with the questions today.

DR. HIATT: Thank you.

Welcome, Dr. Temple.

Open Public Hearing

DR. HIATT: There are no public speakers today, but I think there is a statement that has to be put into the record.

Both the Food and Drug Administration and the public believe in a transparent process for

information gathering and decisionmaking. To ensure such transparency at the open public hearing session of the Advisory Committee meeting, the FDA believes that it is important to understand the context of an individual's presentation.

For this reason, FDA encourages you, the open public hearing speaker, at the beginning of your written or oral statement to advise the committee of any financial relationship that you may have with the sponsor, its product, and, if known, its direct competitors.

For example, this financial information may include the sponsor's payment for your travel, lodging, or other expenses in connection with your attendance at the meeting.

Likewise, the FDA encourages you at the beginning of your statement to advise the committee if you do not have any financial relationships. If you choose not to address this issue of financial relationships at the beginning of your statement, it will not preclude you from speaking.

So, are there any public speakers today?

[No response.]

DR. HIATT: If not, I think we will move on.

First, we will begin with some overview presentations from the sponsor.

Sponsor Presentation

Cellegy Pharmaceuticals, Inc.

Overview of Phase 1 and Phase 3 Studies

DR. AZARNOFF: Good morning, ladies and gentlemen. I am Dan Azarnoff. I am an internist with a subspecialty in clinical pharmacology.

[Slide.]

What I would like to do today is to present some data as to our purpose here, which is to review the Cellegesic clinical trials for which there were three Phase 3 studies, which we have for simplicity called Studies 1, 2, and 3.

If you have read the briefing documents for the FDA and for Cellegy, you will see that there are clearly some disagreements on how the data should be looked at, and what I would like to do is to give you some inkling of how we believe

the data is correctly analyzed and to answer any questions or information that the committee wishes to ask of us.

[Slide.]

The agenda is to discuss the three clinical trials. A colorectal surgeon will tell you about anal fissures. There will be an overview and regulatory history of the trials. Safety will be discussed, a very important discussion of statistical methods and analyses, a risk/benefit assessment, and finally, a summary and conclusions.

[Slide.]

A chronic anal fissure is a tear in the lining, or anoderm, of the terminal anal canal. The symptoms are twofold, one, severe, often debilitating pain, which is in part secondary to increased tone or spasm of the internal anal sphincter, and the other is bleeding.

[Slide.]

Now, Cellegy has developed Cellegesic nitroglycerin ointment 0.4 percent for the acceleration of the relief of pain associated with

a chronic anal fissure.

Efficacy is based on evidence from one Phase 1 trial, which was a bioavailability study, and three Phase 3 studies.

At the present time, there is no medical treatment specifically approved for chronic anal fissure pain in the United States, and I will allow my colorectal surgeon colleagues to tell you about postoperative complications.

[Slide.]

As you are well aware, being cardiologists and frequently using nitroglycerin, that it is converted to 1,2 and 1,3 glycerol dinitrates and to a neurotransmitter nitric oxide. Nitric oxide not only relaxes the vascular smooth muscle, but also the internal anal sphincter, which is smooth muscle, and it does this without causing incontinence.

[Slide.]

The internal anal sphincter since nitroglycerin to the cardiologist is associated with tolerance, interesting enough, in the internal

anal sphincter, nitroglycerin does not cause tolerance, and this has been demonstrated both in experimental animals by Wang and Fung in rats and also by Ciccaglione in humans in which using internal anal pressure as a surrogate for internal anal sphincter pressure, that continuous administration of nitroglycerin, both 0.2 and 2 percent to the anal canal continuously for 12 weeks, measuring pressure before the study and after, had no significant change.

We did one bioavailability study, the mean absolute bioavailability of 375 mg of Cellegesic ointment 0.2 percent. At the time we did this study, we didn't realize that 0.4 percent was going to be the final concentration of the product. That is 0.75 mg of nitroglycerin.

If you look at studies with the nitroglycerin patch, there is a linear relationship between 5, 10, and 15 mg to the blood levels, and therefore, if you just double that to 1.5 mg, you would get twice as much, but the bioavailability, approximately 55 percent.

[Slide.]

We did a dose response in Study 1 and Study 2. There were multiple doses, three doses in Study 1, two doses in Study 2, and if you look at the red line, which is the ointment, you can see that the slope of that curve compared to the placebo, which is the blue line, for 21 days, the dose by day interaction is statistically significant, and similarly, the smallest effective dose is 0.4 percent.

Now, I should make an important point in my opinion, and that is that looking at dose, do not look at the percentage of the ointment, because of you will see that in various studies, there will be different percentages of ointment.

The actual dose of nitroglycerin is the amount of ointment which is applied, multiplied by that percentage in the ointment.

[Slide.]

Now, the incidence of anal fissure, there are currently estimated the incidence to be 765,000 patients in the United States. The visits for that

disorder were over a million, but the important thing is that physicians did use extemporaneously compounded nitroglycerin for treating those patients with an anal fissure.

[Slide.]

So, let's look at the quality of extemporaneously compounded nitroglycerin, which in our opinion, in a study which is about to be published in Diseases of Colon and Rectum, it is very poor.

In looking at prescriptions filled at 24 retail pharmacies in the United States, which was then sent to an independent laboratory for analysis, you can see that if you look at potency, which is required by the USP to be between 90 and 110 percent of the stated label, and if you look at the content uniformity which would be up to no more than 6 percent in the random standard deviation, you can see just looking at these values, there is a significant one above the potency level, and here is one very low.

So, the quality of this material

extemporaneously compounded by local pharmacies is 46 percent outside of the specifications of the USP.

I would like turn the podium over to Dr. Michael Abel, a colorectal surgeon in practice in San Francisco, and on the faculty of the University of California at San Francisco Medical Center.

Pathophysiology of Anal Fissures and Clinical Aspects of Diagnosis and Treatment

DR. ABEL: Good morning. I am Michael Abel. I practice colorectal surgery in San Francisco in private practice and the clinical faculty at the University of California, San Francisco.

[Slide.]

This is a picture of the distal rectum and the anal canal. We will be talking a lot about the anoderm. This is the anal canal which is covered by normal skin, and this is where fissures occur, and the reason they hurt is because this is skin. You will also hear about the internal sphincter, which lines the distal rectum and the anal canal, and

this is the voluntary sphincter mechanism.

[Slide.]

As I mentioned, a fissure is a tear in the lining of the anal canal. The incidence is approximately the same in men and women. People with fissures tend to be on the younger side, 20s, 30s, and 40s most of the time.

The etiology is uncertain although probably spasm of the internal sphincter predisposes or sets the stage for fissures to occur, which may occur in an individual who has constipation or perhaps diarrhea.

Most of the fissures occur in the posterior midline, and there is some question about the adequacy of the blood flow to this area, which perhaps also predisposes to the formation of fissures.

[Slide.]

It is pain that brings patients to my office, and it is relief of this pain that they seek.

[Slide.]

A chronic fissure has scar tissue at the outside opening of the anal canal, called a sentinel tag or pile, and on the inside there is also scar tissue, which is called hypertrophied or large papilla. Often these fissures have an indurated edge and many times you see the exposed white internal sphincter.

[Slide.]

This picture doesn't project very well, but it shows a very small, few millimeter crack or rip in the lining of the anal canal, and this is what causes all the misery for these patients.

[Slide.]

Certainly, traditional treatment, which includes sitz baths, taking of fiber, stool softeners, increased water consumption, anti-inflammatory agents, and perhaps topical local anesthetics works.

Surgery is also an option. The procedure is called a lateral internal anal sphincterotomy, would divide the lowest portion of the involuntary muscle to relieve the spasm. The procedure works,

but it is reported to have complications associated with it, primarily sphincter impairment up to 35 percent of the cases.

Of course, when you do surgery, the cost of healthcare increases significantly compared to medical treatment. As far as medical therapy goes, in addition to the traditional treatment of fiber and water, we often prescribe topical ointment, which is cortisone, cortisone ointment, although there is really no evidence that cortisone ointment does anything in healing anal fissures.

[Slide.]

Nitroglycerin ointment, of course, is an option for us in the treatment armamentarium. Nitroglycerin is metabolized to nitric oxide, which relaxes the internal sphincter, and this, in turn, allows the fissure to heal.

[Slide.]

Nitroglycerin has been recommended by the American Gastroenterological Association, AGA, and my society, the American Society of Colorectal Surgeons through the Standards in Practice Task

Force.

In fact, they gave it a Class I, Grade A recommendation, and to make sure I quote them correctly, the source of evidence for a Class I recommendation is meta-analysis of multiple well-designed controlled studies, randomized trials with low false positive, low false negative errors, high power.

We were introduced in our practice to nitroglycerin ointment in 1998, when our practice participated in the very first study in using nitroglycerin. We have been using the compound ever since.

Only two pharmacies make this formula in San Francisco, so it is clear that in order for our patients to have broader access and perhaps a more accurate and predictable preparation, I think it would be desirable to have a drug available on the market, such as Cellegesic's product.

Thank you very much.

The next speaker is Dr. Azarnoff.

DR. HIATT: We have plenty of time and we

usually discuss these at the end, but if anyone has any questions for Dr. Abel, we might ask them now.

One question I have with the natural history, we noticed in the information provided that pain relief seems to be part of the natural history of these patients.

Can you tell us how often is it recurrent, you know, what happens to these patients after, say, 56 days, kind of the extent of the follow-up in these studies, how much recurrent problem is there?

DR. ABEL: I will tell you about my practice. When I see patients with a fissure, they complain usually of pain, which is often severe. It prevents them from going to work, it prevents them to go to the bathroom to have a bowel movement, and they talk about spotty, brighter rectal bleeding. This is the usual and typical presentation.

Examination is a challenge because they hurt, and there is a lot of sphincter spasm, but once you make the diagnosis, prescribe whatever

treatment you do, and in my case, I prescribe nitroglycerin ointment for every one of my patients prior to considering surgical treatment.

I usually follow them in two weeks, follow-up at two weeks, because what I have observed, that usually, in about four or five days, they start to notice an improvement if they use nitroglycerin, and by two weeks, they reach a satisfactory level.

If the exam is unremarkable, and I am not looking for healing, I am primarily looking for pain relief, if they do well, I do not see them back for follow-up. I tell them to return only if they have problems.

Yes, there are a group of patients who have recurrent symptoms, and the recurrence can occur months and years later, and I see the patients in the office when they have recurrences, so I do not have a long-term follow-up recommendation unless they remain symptomatic.

DR. HIATT: So, healing is not a prerequisite for pain relief.

DR. ABEL: Healing is not a prerequisite for pain relief, that is correct.

DR. HIATT: Do you know if medical therapies, in general, promote healing?

DR. ABEL: Well, I think time in probably any kind of therapy that you use ultimately leads to healing, and I think no matter what you use, whether that's a conservative treatment, nitroglycerin, if you give it time, I think they eventually will heal, most of them, not all of them, but most of them.

DR. PICKERING: Could you tell us a bit more about the basis for the recommendations by the professional societies for the use of nitroglycerin, the evidence on which they were based? Did that include the studies we are reviewing here, or are there other studies?

DR. ABEL: No, to the best of my knowledge, it does not review this study. It looks at published reports, not industry studies. This is a committee of the American Society of Colorectal surgeons that reviews various procedures

and makes an assessment and a recommendation through its publication in Diseases of Colon and Rectum.

DR. PICKERING: What was the level of recommendation again?

DR. ABEL: It was a Level 1 recommendation.

DR. HIATT: 1A, I think I heard.

DR. ABEL: Yes.

DR. LINCOFF: In the briefing materials, there have been presented some data of reduction in the need for surgery historically in countries that use this.

During the time period--and I would be interested in what that time period is in your own practice--have you seen less frequent need for surgery, or have data for that, and also, are there any published data within the organizations to suggest that that is the case and that there aren't other changes in practice that might be associated with that?

DR. ABEL: You will see very nice data

later by Dr. Lund, which shows the UK experience. From my personal experience, about 80 percent of the patients that I see as a surgeon, remember these people are already filtered through the process, I am successfully managing conservatively and with topical nitroglycerin, so I operate only about 20 percent of the people with fissures.

DR. HARRINGTON: I am trying to understand a little bit more of the demographics of the disease. You mentioned that these are mainly younger patients, less than the age I think you said of 40.

Are there other comorbidities that we might be interested in that track with the disease, diabetes, anything else that might be important to us to understand?

DR. ABEL: In general, no, they tend to be healthy individuals without any comorbidities, and again you will see statistics that shows exactly what the age breakdown is later on in the presentation.

DR. HARRINGTON: My other question is

nitroglycerin predominantly, as you say, is working here as a vasodilator. Have other vasodilators been tried in this disease for its treatment?

DR. ABEL: I have not.

DR. HARRINGTON: Do you know of any of that in the literature?

DR. ABEL: Yes, I think there is literature concerning it. I have not tried it.

DR. FLACK: How effective are oral analgesics for treatment of this condition? Does this condition respond very much at all to orals?

DR. ABEL: Again, I don't want to preempt the presentations later. You will see some numbers that deal specifically with acetaminophen as a drug that was used to manage the headache that may be associated with the treatment with nitroglycerin.

My recommendation is to use anti-inflammatory agents now. I have no proof as to how well or how effectively they work in these patients.

DR. FLACK: In your clinical experience, what would you say? You have given us your

clinical experience in observing in an uncontrolled way, the treatment with nitroglycerin, so in that same clinical experience, what about oral analgesics?

DR. ABEL: I don't prescribe oral analgesics as a routine. I never prescribe narcotics, for example, for patients. I only recommend over-the-counters, and of the over-the-counters, in my experience, anti-inflammatory medications work the best. I do not prescribe acetaminophen in particular.

DR. FLACK: So, you have no opinion really on whether they are effective or not.

DR. ABEL: I am not certain how effective they are, that is correct.

DR. TEERLINK: This is a follow-up on Dr. Lincoff's question. Could you explain or describe a bit your decisionmaking process in terms of how you decide to take a patient to surgery?

DR. ABEL: Prior to our participation in the study, by the time patients came to our office, they usually have gone through the conservative

treatment regimen and a cortisone ointment, and we probably operated about 60, 70 percent of the patients, which was one of the treatment options available.

When I participated and our practice participated in the study, it was pretty obvious that something is happening which is allowing us to treat patients conservatively. What happened was the availability of nitroglycerin ointment.

So, when I see these patients, I always place them on nitroglycerin as the first line of therapy, and I only operate on patients who are nitroglycerin treatment failures.

What I like about the drug, again clinical experience based on clinical observation, is that if a patient responds to the drug, usually within the first two weeks, the chances are pretty good that they will respond and they will not return to my office.

If they do not, I often recommend surgery for these patients, because I am quite certain that they are going to have ongoing problems with

fissures.

DR. TEERLINK: So, how long do you usually wait until you just make the decision--

DR. ABEL: I usually wait no less than four and six weeks after I see them, and it depends on the severity of the pain that I make the decision as to when to proceed with surgery.

DR. TEERLINK: So, a difference in pain at 28 days is kind of what would help you decide whether you go to surgery or not?

DR. ABEL: Actually, I see it sooner, usually, about two weeks, and the reason I mention it, because I bring them back for a follow-up visit at about two weeks.

DR. TEERLINK: And from your experience, you would suggest that there is like a 66 percent event reduction in terms of need for surgery based on your clinical experience.

DR. ABEL: Significant. I am a conservative surgeon, and I must say in our practice, we have five colorectal surgeons. Four use nitroglycerin. The fifth one still prefers

surgery.

DR. TEERLINK: Thank you.

DR. LINCOFF: I am sorry, one last follow-up on that. So, as a conservative surgeon, prior to the availability of nitroglycerin, had you ever asked patients let's see how you do in two to four weeks, you know, with continued conservative therapy, looking again at the decline spontaneously in the pain scales over that time period?

DR. ABEL: I usually outline the options, and when I outline the options to patients--and even now when I prescribe nitroglycerin, I always discuss surgical option with the patients--almost always patients select non-operative approach, which is, quite frankly, my preference, as well.

Once in a while, an individual who says no, I would like to be fixed, if you will, as quickly as possible, because I don't want to deal with the pain, are the ones I operate on relatively early.

Prior to nitroglycerin, again, I tried conservative treatment, I gave them cortisone

preparations. These are the patients who usually came back a lot sooner saying, hey, it's not working, I want something else done.

DR. PICKERING: Have other ointments with local analgesics, such as lidocaine, been tried for this condition?

DR. ABEL: Usually, patients pick up topical anesthetics on their own, trying to alleviate the pain. In my practices, I don't like topical anesthetic, I don't like topical lidocaine, they tend to sensitize the skin, give them skin irritation, so I do not use it routinely as a treatment regimen.

Patients often try anything and everything that is available over the counter before they come in and see me.

DR. FLACK: By the time someone gets to you for an anal fissure, painful, how long have they usually suffered, and what you would expect sort of an average or median duration of a painful episode to be for the patient?

DR. ABEL: It varies a great deal

depending on the referring physician. Some send them early, some send them later. It is certainly weeks and often months before they get to me as a surgeon. If I have to give you an estimate, I would say probably within four to six weeks, the ones who have really intense pain. The ones with mild, moderate pain, probably managed by the primary care physicians.

DR. HARRINGTON: Maybe one more question here. I think I got you right. You said that before you started using nitroglycerin ointment, you were operating on 60 to 70 percent of these patients who were refractory in your experience before they got to you, and after you started using it, you are now operating on about 20 percent.

How is the tolerability of the nitroglycerin ointment in your practice? I mean that would suggest a pretty profound effect, 70 percent to 20.

DR. ABEL: Our experience in participating in the study, that really hit us, is there is something there that we should try perhaps before

we recommend surgery, because surgery, as you heard, is not without complications.

Even though our experience is not as high as the published reports from the Mayo Clinic and the University of Minnesota, as far as sphincter control problems, it still can be a problem, and if it happens in a 20-, 30-year-old individual, or an older one, for that matter, it is of significant concern and a quality of life issue.

I am sorry, I lost my train of thought as I was trying to answer your question.

DR. HARRINGTON: The tolerability of the nitroglycerin ointment.

DR. ABEL: Yes. When I talk to my patients and prescribe nitroglycerin, I always mention headache as a possible side effect. I tell my patients that severe headache is very rare, but a mild headache which usually goes away in a short time or responds to over-the-counter analgesic or anti-inflammatory medications, can work.

Very few patients stop using the nitroglycerin even if they have mild headache,

which is interesting, which what it tells me is that the pain from the anal fissure is significant enough that they continue with the treatment and tolerate mild headache if they have. It is not very often that the patient stops.

I have had individuals, however, come back and say, you know, I have had the most pounding, God-awful headache that you can experience. I always review how they apply the medication, and that is one of the reasons I have trouble with the compounded formula. I think depending on where the drug is layered in the container that they get, I am quite sure that what they administer on occasion is substantially greater than what they should as far as the dose goes, and on occasion they administer a lot less than they should.

One patient came back with a container that actually had a 2 percent nitroglycerin preparation in it, and that individual had substantial headache as you can imagine. So, that is one of the reasons why I would like to see something that has a predictable dose, and when I

prescribe the drug, I would like to know what the patient gets.

DR. HIATT: So, you are using nitroglycerin, and you are using what kind of formulation?

DR. ABEL: A compounded formulation. Two pharmacies in San Francisco are mixing the product.

DR. HIATT: Thanks for clarifying that.

Any other questions from the committee? Why don't we move on to the next presentation.

Overview of Studies and Regulatory History

DR. AZARNOFF: What I would like to do now is provide an overview of the studies and regulatory history of the three clinical trials.

[Slide.]

If you have read the briefing documents, clearly, there are FDA issues that relate to effect size, and I should point out that in the last meeting in which we discussed the protocol for Phase 3, we were given a choice by the FDA of effect size or statistical significance as the basis for approval, and we chose statistical

significance, not effect size.

Dr. Gibbons will show you the data which indicates that confounding by dropouts, headaches, and acetaminophen does not occur--does not occur.

I have provided you the data on dose response, and there will be an issue about a quadratic term which we can talk about.

[Slide.]

The primary outcome measure in all three studies or in the last two studies was the rate of change in the 24-hour average pain intensity, and this was recorded daily in a VAS scale in a diary that was provided to the subjects.

I will show you a picture of this because it is of some consequence how we accumulated the data. It was not the difference between active and placebo at any one time point. It is the rate of change, not the difference at any one time point, between active drug and placebo.

[Slide.]

The secondary outcomes measures were the rate of change in defecation pain, again recorded

daily in a Visual Analogue Scale in a diary.

[Slide.]

The methods of analysis were a mixed-effects regression model using all of the available data. Patients recorded their pain every day, and each bit of that data is used in the mixed-effects regression model.

Therefore a statistically significant result is evidence of acceleration of pain relief provided by nitroglycerin ointment over the placebo.

[Slide.]

Here are the three studies. In Study 1, and all three studies, were well controlled in the sense of double- blind, randomized, parallel groups, and placebo controlled.

The entry criteria in the first study were anal pain or bleeding for 30 days and a fissure on physical examination.

The first study had eight arms. It was a dose-response study in which we used 0.375, 0.75, and 1.5 mg of nitroglycerin applied intra-anally

twice a day and three times a day for 56 days or until the fissure actually healed.

[Slide.]

This is the way we measured the amount of drug administered, and remember that the individuals who told you they use compounded material, it is supplied in a jar, so how do you quantitate the amount that is used?

You put your finger in and you take out an amount that has been quoted to half the size of a pea or the size of a pea. It is not quantitative.

What we did was supply the drug in a tube with a syringe you attach to it, and if you squeeze the tube gently, the plunger goes back until it has a stop, and when that stops, there are approximately 375 mg of ointment in that syringe, which was then put on your finger and applied.

So, we quantitated the amount of nitroglycerin ointment which was being applied and therefore, we were quantitating the drug.

[Slide.]

The primary endpoint in Study 1 was

complete healing of the fissure that was done by a blinded observer who was qualified to determine whether, in fact, that fissure had healed, and the secondary endpoints, and they were not ad-hoc endpoints as you have been told, they were actually in the protocol as secondary endpoints.

They were the rate of change in the 24-hour average pain intensity, and the rate of change in defecation pain intensity.

[Slide.]

In this study, there were 289 subjects. Healing was approximately the same in both placebo and active group, in other words, there was no significant healing over placebo, but the rate of change in the 24-hour pain intensity for the active group was significantly better than placebo. It was, however, a secondary endpoint.

[Slide.]

We discussed this with the division director, who at that time was Dr. Lipicky, who agreed with us that pain is an acceptable primary endpoint, but he suggested other modifications for

our second trial, that all subjects should use standard care, which in this instance is the use of fiber and sitz baths.

The subjects were to continue clinical trial material for the entire 56 days even if they healed, the pain was reduced or even eliminated. The purpose of the second trial was to confirm statistically significant pain reduction, and he indicated only one additional confirmatory trial would be required.

[Slide.]

We undertook, therefore, a second trial, again well controlled, in which the entry criteria were again anal pain, and this time at least three times per week, and/or bleeding for 30 days, and again a fissure on physical examination.

Since in the first trial, 0.1 percent, that is, 0.375 mg of nitroglycerin was not effective, in the second trial there were only three arms, 0.75, 1.5 mg of nitroglycerin, and since in the first trial b.i.d. and t.i.d. dosing gave the same results, for patient convenience, we

decided to use only b.i.d. dosing in subsequent studies.

The material was applied intra-anally for 56 days irrespective of healing as required, and to provide and be sure subjects were taking standard of care, we actually provided them psyllium, that is, Metamucil, which you may know by the trade name, 3.4 grams to be taken twice a day.

[Slide.]

In this study, the primary endpoint was not healing, but the rate of change in the 24-hour average pain intensity, and that data was analyzed by a mixed-effects regression model. Again, I should point out it is not the difference between active and placebo on any one day.

The secondary endpoints were the rate of change in defecation pain intensity, and healing was also a secondary endpoint.

There were 219 subjects. The rate of change was significant. The defecation pain was also significant, but the healing again was insignificant.

[Slide.]

The mixed-effects regression model in order to obtain significance included a quadratic term. The FDA would not accept this inclusion of the quadratic term since it wasn't prespecified or in the statistical analysis plan, but, in fact, it is a routine part of a mixed-effects linear regression model when the data becomes curvilinear, and, in fact, as has been shown, the Medicines Healthcare Products Regulatory Agency in the United Kingdom assessors accepted and approved Cellegesic for marketing in the UK based only on Studies 1 and 2. Subsequent to that, it has now been approved in 19 other European countries based on all three Studies 1, 2, and 3.

[Slide.]

Following the second study, we had further discussions with the Division, and we agreed under a special protocol assessment on the basis for Study 3. The primary endpoint would be the rate of change in the 24-hour average pain intensity during the first 21 days of treatment, and that was

selected, because for the first 21 days, the response is linear and then subsequently becomes curvilinear as you will see.

Dropouts due to nitroglycerin-induced headache would have their last observation carried forward, which we objected to, but did, in fact, include in the trial rather than the standard mixed-effects regression model.

[Slide.]

Headache, a nitroglycerin-induced headache was defined in the protocol. In Section 9.4.2 in the protocol, you will find a definition of a nitroglycerin headache. I should remind you that under a special protocol assessment, both the sponsor and the agency agreed to the protocol. In that protocol, there was a definition which was agreed to.

The FDA also agreed that three subjects dropped out for nitroglycerin-induced headaches. We believe this is documented in the minutes of a meeting which was held on March 31, 2005.

[Slide.]

Here is the Phase 3 study. It was well controlled. The entry criteria was somewhat different. In addition to the presence of a fissure, the fissure had to have a sentinel pile. A sentinel pile, as you heard from Dr. Abel, is an indication of chronicity of the fissure. But in addition, the subjects on the two days prior to enrollment had to have pain on the VAS scale of at least 35 millimeters.

They similarly had to have moderate or severe defecation pain on a categorical scale on at least one of the two days prior to enrollment.

As I said, the physical exam required the presence of a sentinel pile.

We settled on two arms, 1.5 mg of nitroglycerin, that is the 0.4 percent ointment, which will be what you will usually hear in the rest of the discussion, applied intra-anally every 12 hours for 56 days irrespective of healing. Fiber and sitz baths was allowed as standard care only if it was used in the previous week prior to entry.

[Slide.]

The results, as you see here of the trial, I am not going to go into detail, but even using LOCF, we believe the trial was statistically significant at less that 0.05, and if you use the standard LOCF, it is less than 0.0309 as a p-value.

Healing against was not statistically significant, but similarly, the 21-day and the 56-day pain for fecal pain, defecation pain was also significant.

[Slide.]

Now, I show you this to show you how we accumulated the data and how we believe we have accurately determined whether a subject had a nitroglycerin headache.

This is a page out of the diary given to the subjects. They were given a single page in this diary for every day in the study. They were to record their 24-hour average pain intensity, which you see here. They also were to record the defecation pain on the next line, but they also were to record whether, in fact, they had a sitz

bath that day and also whether they had a headache.

[Slide.]

Also, please note that they were to record the time at which they applied the drug in the morning and the time at which they applied the drug in the evening, and if they had a headache, as you will see, they were given another diary page in which they were to record when the headache started and when the headache stopped, but the important point is from this data, you can now detect whether the headache occurred within 30 minutes of administering the drug, which was the definition of a nitroglycerin-induced headache that was accepted in the protocol.

[Slide.]

So, we can conclude, then, that Study 1 provided evidence that relief of pain, not healing, was the appropriate primary endpoint; that Cellegesic nitroglycerin ointment rate of change in the 24-hour pain relief was significantly better than placebo.

Analysis of Study 2 revealed that the pain

relief was linear for the first 21 days of treatment with Cellegesic nitroglycerin ointment and curvilinear thereafter, but by using linear effects with a quadratic term, it was statistically significant.

In Study 3, nitroglycerin ointment, the rate of change in the 24-hour pain intensity was significantly better than placebo over the first 21 days of treatment at less than 0.05 even with using the last observation carried forward, and Dr. Gibbons will discuss that at greater length. Without imputation, it was less than 0.0309.

[Slide.]

So, in conclusion, Studies 1, 2, and 3 were reanalyzed using the same method for the data up to 21 days, and that data was significantly better both for the first study and the second study, and it was also much better when all three studies were combined.

Safety

[Slide.]

Let me move briefly to safety. This was

not a big issue, but we can discuss it briefly.

[Slide.]

The data for safety came from one Phase 1 and three Phase 3 studies that were well controlled. They came from three different doses of nitroglycerin applied twice a day and three times a day up to 56 percent.

[Slide.]

You see here the demographics of the baseline characteristics. Males and females in the study were approximately the same. The study did include mostly Caucasians and mostly individuals in the younger age group, as you see here.

[Slide.]

Adherence was measured by weighing the tubes before they were distributed to the subjects and when they were returned, unbeknownst to the subject that they were being weighed, and the average amount expressed per day was determined, and the amount was approximately the same, 100 percent on average in both the placebo and active group, although we did decide that between 70 and

130 percent of the amount to be used was a satisfactory amount.

[Slide.]

The subject disposition using all of the studies in the trial was 721. 592, 82 percent, completed the trial. If we look specifically at those individuals who were taking the 4 percent b.i.d., which is the requested approval dosage, 9.7 percent, 20 subjects dropped out for adverse events, whereas, in the placebo group, only 2.8 percent, approximately 3 times more.

[Slide.]

Here, we see the frequently reported incidence of treatment-emergent adverse events greater than 2 percent, and here, it is quite obvious that headaches do occur, they occur fairly commonly in our study, but the other interesting thing is we believe we may have sensitized people to the incidence of headaches, because if you look at the placebo group, they had almost 39 percent headaches, and that group also was using the same diary in which they were being asked every day

whether they had a headache or not.

If you look at other studies like with the nitroglycerin patch, the incidence of headaches is about 18 percent in the placebo group.

The other things which were greater than 2 percent was dizziness was 4.4 percent, nausea was 5.6 percent, and in the placebo group, diarrhea was slightly greater than the active group.

If you look at the same data in which the investigator considered they were treatment related, the numbers go down slightly, but, in fact, the sort of relationship remains approximately the same.

[Slide.]

If you look at subgroups, some interesting things appear. If you look at headaches in the active group, you can see that females are more likely to complain of headaches than males, that the younger age group are more likely to complain of headaches than the elderly.

If you look in the placebo group, you see the same sort of ratios. If you look now at

nausea, again, females appear to have it more frequently than males, but unlike headaches, the older individuals were more likely to experience nausea, and in the placebo group, the levels are low enough that we really can't see any major differences in subgroups.

[Slide.]

So, let's look at headaches briefly. In the Cellegesic group, they were 71 percent compared to 30 in the placebo group, but the important thing is if we go to the bottom and look at individuals who had a headache in the first week, there was 71 of them, and by the fourth week, although only 65 subjects were left, the incidence of headaches was now 32 percent, certainly suggesting that there was tolerance developing to the headaches and also, at the same time, there was a decrease in the severity of those headaches during that period.

[Slide.]

There were no deaths during the study. There were 10 serious adverse events, 6 of them with Cellegesic, 4 with placebo. The only one that

investigators considered related to the drug was a subject who had a history of migraine headaches, who developed a severe migraine headache on the first day of treatment with a 0.75 mg dose of nitroglycerin, and I should point out that there is documentation in the literature that if you have a history of migraine headaches or chronic headaches, it is very likely that you will develop a nitroglycerin headache or nitric oxide headache, if you will, when administered nitroglycerin.

With the serious adverse events, there are no cardiovascular adverse events considered related to the drug.

Discontinuations in the three studies due to adverse events. In the Cellegesic group, there were 20 subjects who dropped out, 9.7 percent; in the placebo group, 7 subjects, 2.8 percent as I previously showed you.

But an important thing is if you look at the headaches, in the Cellegesic group, 16 of the 276 subjects who dropped out due to headaches, that's 7.8 percent, 9 of the 16 had pain

improvement by the time they dropped out, 9 of the 16 had a VAS scale less than 30, which is mild pain, and 10 of the 16 had either pain improvement or a decrease less than 30 mm, indicating they did not drop out until their anal pain was significantly reduced.

[Slide.]

There were no consistent clinical laboratory hematology or clinical chemistry changes.

[Slide.]

There was an issue brought up by the FDA related to cardiovascular effects. We did measure blood pressure at every visit, and in the first trial, we measured blood pressure at 10 and 20 minutes following administration of the first dose, we looked for a 20-mm drop in diastolic pressure as an indication of an effect.

If you look at this data, you will see that it is the lowest dose group which had the greatest drop in blood pressure, that is the 0.375 mg group, and, in fact, there is no trend even in

blood pressure changes with increasing dose.

[Slide.]

So, in conclusion, the safety of nitroglycerin, as you well know, has been established, and we know what the adverse events are for over a century of use.

The dose of nitroglycerin ointment, that is, 1.5 mg applied intra-anally every 12 hours, is less than that of many of the approved nitroglycerin products currently on the market, like 5, 10, and 15 mg of the patch.

The adverse events observed in the three, Phase 3 trials are consistent with the known adverse events of nitroglycerin. Headache can be managed with mild analgesics has been shown in trial with the patch, and other adverse events should be accounted for and taken care of in appropriate labeling.

I would now like to introduce Dr. Robert Gibbons, who will provide you with an analysis of the methods used in analyzing these data.

DR. HIATT: I wonder if we might just

pause a second here. This committee likes to discuss things, and we can hold this to the end, but it is often more productive to kind of hear the clinical data being presented.

I know there are some biostatistical questions that will come up, but I thought it was okay, we have plenty of time, we are ahead of schedule, if we could go ahead and maybe ask you a few questions about what you have just presented.

Let me just lead off with one. In the first study, you tested multiple doses, and you did b.i.d. and t.i.d. dosing. I tried to look for a dose response in that data, looking at the slope changes, the endpoint, and I was able to kind of go back and look at the total milligrams delivered on a daily basis.

It went from 0.75 at the low to 4.5 mg as the high dose, and then plotted the effect size as the slope across those six doses that were delivered. It was very flat. In fact, the effect size of 0.16 was seen at the second to the lowest dose, was the same effect size at the highest dose,

and the only dose that gave statistical significance was the 0.4 percent given b.i.d., bracketed by insignificant responses below and above that dose.

So, I can't see a dose response here. When you then look at the reviewer's comments later in our document, looking at the meta-analyses of placebo versus 0.2 versus 0.4, there is sort of a trend for a negative effect at 0.2 and a positive effect at 0.4.

So, my first question of many is why is there an absence for dose response in these data?

DR. AZARNOFF: If you look at the data, I presented a dose-response curve from the combined first and second trial. Those were the two trials in which there was more than one dose. The data shows that there, in fact, is a dose response, and the analysis was done by Dr. Gibbons, and if there is further questions regarding that, I will let him explain to you how he analyzed that data.

DR. FLACK: The drop in blood pressure, you took 20 mm of mercury, is that an acute drop,

is that a drop over time that you were reporting, that 20-mm mercury drop? What was the starting point and how far down did you--

DR. AZARNOFF: The blood pressure was measured during each visit. It was measured only once. It was measured as part of the physical exam at each visit, so I can't give you data on what happens with blood pressure over time except in the first study, where we measured the blood pressure in subjects following their first dose at 10 and 20 minutes, and we did not see any markedly decrease in blood pressure.

DR. FLACK: So, you recorded the absolute values, both systolic and diastolic, but just reported it out as a threshold value of more than 20 mm of mercury, because that is a huge drop, and that's more than you get chronically with an antihypertensive drug, and that would not be really reassuring from a safety perspective and probably more for issues of acute changes, systolic pressure would be probably more important than diastolic.

One final question. You made an issue of

the definition of nitroglycerin headaches, yet, the data you showed, almost any way you presented it, you were almost two times as many headaches in the nitroglycerin group as the placebo group, or just under two times as many.

Even if a headache doesn't fit the definition of a nitroglycerin headache, do you still believe it is possible that nitroglycerin is sensitizing people to headaches and that these are really treatment-related headaches, and, if not, what would be a plausible explanation?

DR. AZARNOFF: Why do people drop out from nitroglycerin?

DR. FLACK: No. Why do they get the headaches almost 2 to 1?

DR. AZARNOFF: Why do they get them. The mechanism for a nitroglycerin headache is the conversion of nitroglycerin to nitric oxide, and the nitric oxide releases a substance calcitonin G protein, and that causes the headache. That's the current theory as to a nitroglycerin-induced headache.

DR. FLACK: Okay, but why do people get it? People are getting it almost twice as often in the treatment group as the placebo group, and a big deal was made about the small number of true nitroglycerin headaches, and I am just trying to find out, is there a rationale that you can provide to why people would get it if it wasn't nitroglycerin related.

DR. AZARNOFF: I can't answer your question exactly, but, in fact, if you look at an individual over 8 weeks, I don't know about those of you in the audience, and I only speak for myself, but how many people develop a headache.

If you look in the literature, in the clinical pharmacology literature regarding adverse events, if you take a group of healthy individuals, they will, in fact, develop a headache and list it as a significant adverse event many times during a time when they are not taking any drugs.

So, it is not surprising to me at least that subjects in a trial, where you are asking them, in fact, every day whether they have

developed a headache, that they indicate they have had one.

DR. FLACK: Just one follow-up comment. I agree with you that people will report things, and particularly if you don't wait for them to self-report, if you are asking them, you are going to get a higher rate, but it shouldn't be differential by group unless there is something going on with the treatment. I will admit that this study was probably very difficult to blind given the fact that headache was so common for both the investigator and the patient.

Did you try anything at all with trying to quantify the analgesics they used and their impact, because you used mixed-effects models, which likely put in time dependent covariates?

DR. AZARNOFF: Dr. Gibbons will provide evidence that the analgesics did not have an effect, number one, but number two, I want to reiterate that the definition is in the protocol, and the definition was approved in a special protocol assessment by both the FDA and the

sponsor.

DR. HIATT: I think, as we go around, there will be a statistical presentation and maybe we, in fairness, should hold a lot of our statistical questions until after that.

I also think, Bob, you are going to give an FDA perspective on this, and so I think there is a lot of material and maybe some things we want to go through.

I think what we are trying to do right now is just understand the clinical data you have presented, both efficacy and safety, and I would like to just say, Dr. Flack, I appreciate the safety concern you brought up. Before we are done today, we need to flesh that a little bit more. Certainly the tolerability issues have been well presented.

I realize there is a history of nitroglycerin that goes back many years, but in this particular population, the absence of finding a safety concern doesn't exclude a safety concern, and I think the committee needs to wrestle with

that issue before we are done deliberating.

Tom.

DR. PICKERING: I have a question about the evaluation of pain. We are basically talking about two types of pain. There is the anal pain which was scored on a Visual Analogue Scale, and the headache which used a different scale. I think it was 1, 2, and 3.

My question really is what instructions were the subjects given about how to score their anal pain. I think it said no pain to worst pain imaginable. Can you fill us in a bit more, because the sort of context of the instructions to the subject might have a huge influence on where they actually chose to mark off the score?

DR. AZARNOFF: The instructions to the subjects were that each night before they went to bed, they would complete their diary, and in the diary, they would record an average of the 24-hour pain. That is, an average of their pain, their anal fissure pain. It was very specific, their anal fissure pain over the previous 24 hours, and

to mark that with a hash mark on the Visual Analogue Scale.

The distance from the lefthand end to the hash mark in millimeters is the intensity of the pain. In general, if you divide that scale into thirds, the lower third is mild, the middle third is moderate, and the righthand third is severe pain. The subjects weren't told that.

In regards to headache pain, they were told that they would use a categorical scale, which is mild, moderate, and severe.

DR. PICKERING: Can you tell us how compliant the subjects were? I mean they were asked to do this every day for 56 days. How complete were the actual pain ratings, because it is asking a lot of subjects.

DR. AZARNOFF: Correct. Not everybody obviously completed every single day. I was surprised actually at the amount of data which we did collect. It was very extensive. Most subjects did comply with the requirement to complete the data.

If data were missing, you will hear that in the analysis which we did, that data can be compensated for, and I think it is important that Dr. Gibbons tell you how we handled all that data.

DR. PICKERING: One related question. In one of the FDA figures, which was on page 13, sort of towards the back, it shows that for individual subjects, daily ratings of pain, and there was an absolutely huge variance for some subjects, between nearly 100 and nearly zero on successive days.

Was that sort of typical, I mean if you look at the individual subject's pain ratings, did they fluctuate that much?

DR. AZARNOFF: I think, if you don't mind, that particular data will be discussed, and I think somebody who is going to do it can analyze that for you quite satisfactorily.

DR. HIATT: Why don't we keep going around the room just to make sure everyone addresses their clinical questions. Related to that, if you have a treatment that relieves one form of pain, and perhaps induces another, why would you not use a

global pain index?

DR. AZARNOFF: Once again, please.

DR. HIATT: Obviously, patients come because they have rectal pain, and your primary endpoint is relieve that pain, but if your therapy might induce another kind of pain, why wouldn't you use a more global pain index?

DR. AZARNOFF: There was a discussion about whether or not to use a global pain measurement at one time. I don't remember the reason why, but the sponsor decided not to use that.

DR. HIATT: Why don't we just keep going around the room and just make sure everyone gets their questions addressed.

DR. KOLTUN: Don't forget me on the phone because I have a couple, too, please.

DR. HIATT: We didn't hear that.

DR. KOLTUN: This is Dr. Koltun on the phone, and I have some questions, too. I am just trying not to be forgotten.

DR. HIATT: Since you are on, why don't

you go ahead and ask your questions.

DR. KOLTUN: Okay. Since I am a practicing colorectal surgeon myself, and have some experience with both the disease process and this therapy, a couple of specific questions.

First, were finger cots used in the context of the application of the drug?

DR. AZARNOFF: A pump?

DR. KOLTUN: No, finger cots.

DR. HIATT: Finger cots, how was it applied?

DR. AZARNOFF: The material was supplied in a tube. I showed a picture of that tube and how a material was put into a syringe. The syringe was emptied onto a finger, which was covered by a finger cot, and the instructions were for the subject to gently insert it into his anal canal up to the first knuckle, which is about the area where the fissure would be, and to gently rotate it around the anal canal.

DR. KOLTUN: Thank you. My second question is in regards to the pain assessment,

which obviously is a critical aspect of all this, the pain associated with a fissure most typically is rather severe immediately at the time of bowel movement, and usually lasts for some period of time thereafter, usually on the order of a couple of hours.

So, my understanding is that you determined or you requested one pain assessment on a daily basis in the evening, is that correct?

DR. AZARNOFF: That's correct, one pain assessment in the evening as an average for the entire day.

DR. KOLTUN: Did you take into account the issue that the pain is actually quite variable during the day based on bowel movement and, in fact, could be much more extreme to the patient who had three bowel movements, for example, and three separate episodes of two hours of post-bowel movement pain versus a patient who just has one?

DR. AZARNOFF: Yes. Clearly, the amount or the frequency of defecation would have an effect on the pain. In addition to looking at the average

daily measurements, we also looked at defecation pain, which would account for that type of effect of defecation. We did not compare the defecation versus the average, because we didn't have that kind of data available.

If a subject had a defecation at least once a day, they just recorded they had defecation.

DR. KOLTUN: Now you are confusing me, because I thought you just said that you only requested one pain assessment per day, but now you are saying that you also at times requested a defecation pain assessment?

DR. AZARNOFF: There were two assessments. One was for average daily pain, and the other was for defecation pain every day.

DR. KOLTUN: So, there were two numbers requested on a pain scale every day.

DR. AZARNOFF: Correct.

DR. KOLTUN: So, now I ask you, what about the patient who typically--who is not atypical, and that is the patient who has such severe pain, which is again not atypical, that they don't even have a

bowel movement during the day, and sometimes, because this disease process does predispose the individual to basically constipation, and thereby, you know, they might go one day, two days without a bowel movement and have zero pain scores, but then the next day, or the third day, they have a severe bowel movement that gives them a 10 out of 10 type of pain, and then they will go two more days with zero scores?

I mean how was that issue managed in the context of this pain assessment?

DR. AZARNOFF: Clearly, the pain can vary from day to day, and in the analysis, that is taken care of, because every individual measurement of pain, that is, average pain in 24 hours, is put into the analysis, as you will hear.

Similarly, every time the patient had a defecation, the last one for the day, the pain was measured and was again a secondary pain measurement analysis.

DR. KOLTUN: Are you saying that defecation pain indices were grouped and put into

the same statistical plot as the average pain for the day, or were those segregated and analyzed separately?

DR. AZARNOFF: There were two separate analyses.

DR. HIATT: Defecation pain was a secondary endpoint.

DR. KOLTUN: So, when you talk about the improvement in the rate of pain change, you are talking about both those analyses?

DR. AZARNOFF: We are talking about each one separately.

DR. KOLTUN: The last question is we talked about variability of this compounded material and the variability between pharmacies. Can you comment on the issue of shelf life in that, you know, nitroglycerin, generally speaking, has a certain potency that deteriorates over time?

DR. AZARNOFF: We were not able to do a study to determine whether that variability was related to any type of reaction be it an adverse event, lack of efficiency, or what have you.

DR. KOLTUN: What I am basically asking is how do you know your drug, after 56 days, coming from the same vial, was still the same drug that started at day zero?

DR. AZARNOFF: Would you repeat that, sir?

DR. KOLTUN: Nitroglycerin tends to deteriorate with time.

DR. AZARNOFF: Correct.

DR. KOLTUN: And you are having a study that lasts two months, what was the potency of your drug material at the 56-day point relative to the day zero day point?

DR. AZARNOFF: We determined the stability of the material, and it was satisfactory throughout the study.

DR. KOLTUN: Satisfactory meaning?

DR. AZARNOFF: Meaning it was within USP specifications.

DR. KOLTUN: Okay.

DR. HIATT: Thank you. Why don't we keep going around the room for questions.

DR. WARNER-STEVENSON: I had two points.

One, in terms of the headaches, I certainly agree with Dr. Flack's point. I don't think it is possible to define a nitroglycerin headache as just occurring within 30 minutes of the dose. Certainly, our experience with cardiovascular nitrates is it is very variable when the patient develops the headache in relation to the dose. Admittedly, that is oral administration usually, so it is different.

The other question I had related to the time course of efficacy. It really looks like the effect that we are seeing here is concentrated between 7 days and 21 days, looking at some of the analyses, and depending on which measure you are looking at, it seems like it isn't clear until perhaps 5 to 7 days, and other ones it looks like it is clear earlier.

I would actually like to ask both Dr. Abel and perhaps Dr. Koltun, as well, your experience in terms of when does this drug have its most obvious effect, if is early on, or is it after the spontaneous improvement has occurred?

DR. ABEL: I started to see effect usually about the fourth day, and it peaks at about two weeks.

DR. KOLTUN: If you ask me my opinion in regards to its clinical efficaciousness, I dose the drug differently. I dose the drug in the context of bowel movement, so the effect is, I assess its effect more acutely.

Not atypically, I find that if you apply the drug after a bowel movement and get relief of pain, then, that is a fair indicator that it is doing its job and then will continue to do so.

DR. HIATT: Why don't we just continue around the room. Dr. Teerlink.

DR. TEERLINK: I would second Dr. Stevenson's point about the headache being very variable in timing, and we see it, as well, with topical and even intravenous nitrates where patients won't have headaches, and when you first start the IV nitrates, and then even hours later develop a headache. So, I think we are all pretty much agreed on that point.

It seems like you here described the decrease in headache later on due to tolerance to the effects of the nitroglycerin ointment, so there is clearly a time-dependent nature to that assay.

In terms of the nausea and the dizziness, which are also probably related to its hypotensive effects, did those also have a similar time course in terms of reducing over time, or were they fairly evenly distributed over time? There is a 6-fold increase in those events, as well.

DR. AZARNOFF: The headache decrease?

DR. TEERLINK: In terms of the time course of the events of nausea and dizziness, did those also decrease over time?

DR. AZARNOFF: I didn't look at that, sir.

DR. TEERLINK: So, on the one hand, we have that there is this tolerance to these hypotensive effects, but then you also suggested there is not tolerance to the local effects. Is that what you are suggesting?

DR. AZARNOFF: That is correct.

DR. TEERLINK: The other issue is when we

try to evaluate safety, we are looking at potential at-risk populations, and you suggested that the elderly had a higher incidence of nausea, and we have actually seen that, as well, with intravenous nitrates in some of the studies, that there is a slight increase among the elderly, it is unclear why although it may be due to phenomenon from the mesenteric vasculature.

Is there a concern, or do you feel you have studied enough elderly patients to say that it is safe in the elderly?

The second issue is also in African-Americans, if there is an increased sensitivity to nitrates, do you think you have evaluated enough African-Americans to suggest that it is safe in those patients, as well, given that we really don't have any safety data in terms of its acute effects?

DR. AZARNOFF: Most of the subjects who have an anal fissure are young, healthy, and do not, in general, have cardiovascular disease, although many young people do, I understand that,

but it is the younger population who develop this disorder.

The elderly do, in fact, develop nausea slightly more than the younger. It was quite low, a small number. They did not find it sufficient to drop out in general.

DR. TEERLINK: It was a low number, but there were relatively few elderly patients, so among the elderly patients, it would seem to have been a high percentage.

DR. AZARNOFF: It was above 2 percent, it was about 5 percent.

DR. HARRINGTON: I have a couple of questions on the operations of the study. Following up on what Dr. Flack had raised, the concern about your ability to protect the blinding of the study. I want to make sure that I have this right in terms of the transcription of the Visual Scale.

The patients recorded their diaries and brought it back to the study nurses, who then transcribed that information into the case record,

is that correct?

DR. AZARNOFF: Correct.

DR. HARRINGTON: How did you protect the blinding in that situation, were these the same study coordinators that asked about headache, the frequency of tolerability of the drug, et cetera, and did you do any monitoring to assure that there was accurate transcription and no bias being introduced with that process?

DR. AZARNOFF: All of the sites were monitored both by clinical research organization monitors, and in some instance, by Cellegy's own clinical research associates. Several of the sites were audited by outside auditors, as well as by FDA.

DR. HARRINGTON: Did you have data that would suggest that there was good transcription rates of the patient diaries?

DR. AZARNOFF: The audits which I saw did not suggest that there were difficulties.

DR. HARRINGTON: My second question about the operations is I am trying to understand, in

Study 2, you required Metamucil and sitz baths as part of background therapy, but then in Study 3, you didn't require this.

Number one, why the change, and number two, can you provide some data that would tell us that the use of those standard therapies was balanced between the groups?

DR. AZARNOFF: It was used in the second trial because Dr. Lipicky suggested we compare it to standard of care. Following the second trial, and as well as in the first trial, Dr. Gibbons did an analysis indicating that the use of sitz baths and fiber did not influence the results. In fact, in regard to sitz baths, the placebo group took slightly more, although they weren't statistically different, sitz baths than did the active group.

In regard to the third trial, we decided that it was only necessary if people were already doing it, that they should continue standard of care. In looking at the results, remember that the active moiety, the nitroglycerin group, is being compared, not only to the placebo, but to standard

of care. They are being given fiber if they are taking it, and in the second study, everybody took it, and they are taking sitz baths, both of which have some effect on the discomfort.

DR. HIATT: Just to clarify, that is not a comparison. That is background therapy, they all get the same.

DR. AZARNOFF: Sir?

DR. HIATT: Just to clarify what you just said, that there is only one comparison. It's active versus placebo, that you are not comparing your treatment to background therapy.

DR. AZARNOFF: Those were included in the placebo group, as well as in the active group.

DR. HIATT: That is the point.

DR. HARRINGTON: That is exactly, Bill, what I am trying to get at, is was the use of that background therapy balanced between the two groups, so that we can be assured that the primary comparison that we are interested in is a valid one.

DR. AZARNOFF: As I pointed out, Dr.

Gibbons has done analysis which indicates there is no significant difference based on the availability or use of fiber.

DR. HARRINGTON: My final question, following up on something that John asked, with the dizziness, 9 patients, over 4 percent dizziness, was that associated with any bad outcome, pre-syncope, syncope, anything that was reported that would raise caution?

DR. AZARNOFF: Not that I am aware of, sir.

DR. HARRINGTON: Was that specifically asked on the case record?

DR. AZARNOFF: I can't answer what the investigator did.

DR. HARRINGTON: Symptoms of hypotension, particularly syncope or pre-syncope?

DR. AZARNOFF: All I know is what the investigators reported.

DR. HIATT: As we go around, just to clarify one of Dr. Harrington's questions about blinding, on Table 19 in our background, page 48,

it appears that from your analysis, both the subjects and the investigators were more likely to guess they were on active, when they were on active, were more likely to guess they were on placebo when they were on placebo, so one of my big questions is whether the blinding really was maintained with a drug that has vasoactive properties.

DR. AZARNOFF: You are correct in that there were slightly more, but they weren't different. Analysis indicated there were no differences in that. Frequently, investigators thought they were also on placebo when they were on active.

DR. HIATT: I am just looking at your numbers. About 72 percent of the subjects felt they were on active, and in those on placebo, 65 percent. I didn't see a statistical analysis of this, but the gradient favors the placebo, who were on placebo, as well. I am just pointing that out.

Why don't we keep going around the room.

DR. LINCOFF: I would like to focus a

couple questions on the headache. Again, I echo the previous comments that I don't think timing is a valuable criteria to determine nitroglycerin-related headache, and I would be sort of interested in where the source of that came from.

But more importantly, because the p-value seems to be so critically dependent upon the exclusion of one patient or not based upon the cause of their headache, in the materials that we have in the statistical review from the FDA, we don't have the protocol per se, but the materials quote, the initial protocol saying that there would be exclusion for headache, and then the data plan, which subsequently states for nitroglycerin-related headaches.

So, you have said that it's prospectively defined of whether or not it was nitroglycerin related or not, and we can argue whether that definition is meaningful or not, but can you please clarify if the initial analysis plan, not the one that came with the data, but the initial analysis

plan prior to the data being available did or did not specify that it would be nitroglycerin-related headache as compared to any headache excluded?

DR. AZARNOFF: It was prospectively identified that nitroglycerin-induced headaches would be those individuals who were analyzed as part of the dropouts who had had their last observation carried forward. That was done prospectively.

If you look in the protocol, you will have a definition, as I indicated, for a nitroglycerin headache. If you look in the previous two protocols, you will not find a definition for a nitroglycerin headache. The reason it was put in the third protocol was so that we could determine which subjects dropped out for a nitroglycerin headache.

DR. LINCOFF: Perhaps we can get some clarification later then regarding what is in the statistical analysis from the FDA.

My second question related to that, and it speaks to the question of whether or not analgesia

use had any influence on the pain, but I see in the diaries that you collected whether or not analgesia was used.

DR. AZARNOFF: Whether or not what, sir?

DR. LINCOFF: An analgesic, a medication was used for headache. Is there anywhere in the case report form that actually collects how much? I mean there is a difference if a patient takes one dose of Tylenol or acetaminophen as compared to taking it, you know, every six hours throughout the day given that your headache may be short term, but the analysis of pain for the primary endpoint is an average over the 24-hour period.

So, anywhere did you collect and in the analyses represented elsewhere, the influence of analgesia, is that related at all to how much in the way of pain medications were used?

DR. AZARNOFF: Each time the subject took acetaminophen, they were required to record it in their diary. Secondly, they were only allowed to take 8 doses of acetaminophen over the first 21 days of treatment, which is the duration of the

primary endpoint.

So, even if they took them three times for 8 doses, they could only take it for 3 of the 21 days, if they took it once a day, they could only take it 8 of the 21 days. I don't actually have the actual data, but they obviously can't take large numbers for long periods of time.

DR. HIATT: Dr. Stockbridge or Dr. Temple, I know, Bob, you had your light on. Did you want to maybe say a few words?

DR. TEMPLE: I just had a couple of things. It sounds like with respect to headaches, whether you rate them as nitroglycerin related or not, there were more of both kinds in the treatment group, which sort of makes you think the treatment group must have done it even it if wasn't a formal--I mean I think that's the point Dr. Flack was making.

This isn't really a criticism of the study, because a lot of people use diaries, but there is a fair amount of data that people don't always fill out diaries contemporaneously at the

exact moment, the so-called parking lot effect, and stuff like that.

I have never worried too much about that, because I think those are all biases toward the null, they probably don't enhance or it makes the diaries more like a global or a number of other things, but it is sort of worth keeping that in mind. That is one of the reasons electronic diaries are becoming more attractive.

Then, I just wanted to comment on your question, why didn't you combine total pain, and that is an issue that comes up a lot. It is generally the question of should you combine the good things a drug does with the bad things and get an overall effect.

My bias is that you should not do that because they are not really the same thing and people will have preferences about whether their pain is lower or upper, and the pattern during the day is different, so I mean this is a simmering controversy, but on the whole, teasing them apart seems preferable to me, because they might occur in

different people and you should know which is which.

You can certainly subtract them in your head and calculate whether there is a net benefit or not, and stuff like that, but combining them in a single way doesn't seem like the best thing to me.

DR. HIATT: Although we combine often, you know, a bleeding event and a primary endpoint with an ischemic event, so we often combine the--sometimes you have to bundle these things and sometimes you don't, and I think we will wrestle with that as we go through the data.

DR. TEMPLE: Right, where they are of comparable weight and undermine the benefit in an unequivocal way like a stroke that bleeds and a stroke that doesn't, right, but it is not common to subtract symptoms from some other benefit, and I think the reason is that people value them differently, and you may want to know how each person, they may want to choose which they want, the benefit or the risk, and the pain up top or the

pain down below.

DR. HIATT: To your point, the characteristics that we are discussing may be different, temporally different, respond differently to analgesics.

John.

DR. TEERLINK: So, in that context, though, do you think it would be reasonable to make suggestions, that if you are going to look at a kind of cause specific pain or a cause specific symptom, that in addition, you request or there be some assessment of the patient's global assessment, so we at least have the information, or the patient has the information to make that decision between the two?

Otherwise, we are left with, you know, a 2-fold-plus increase in headache versus--well, I guess we will be told what it's versus later on as we hear more about the statistical analysis.

DR. TEMPLE: This goes to a larger question that I will touch on briefly later and that we are thinking about. We tend to look at mean

results, and mean results are not what happens to individual patients. What you sort of want to know is how many patients are better on this scale down below and how many are worse up top.

Trials are not generally designed that way although it is often possible to break the data down in that way, and you can always get a distribution of results, and one of the things we are actively thinking about is whether having achieved a significant benefit on the mean, we ought to routinely display the distribution of results.

We actually have done that. If you look at--not that any of you need this--but if you look at the Alzheimer's drugs, they all show a cumulative distribution score. You could do a cumulative distribution, or you could do a bell-shaped curve, but it really does show the distribution of results.

The results are completely predictable in some sense. Things that have an average benefit always shift the curve a little bit to the side

that you would predict, but it does give you a somewhat sort of visual image about what the distribution of individual results are, and it is sort of attractive, and we are actively thinking about whether one ought to do that.

In this case, you might be able to think about individuals and how many people had rectal pain, anal pain, that they felt good about, and didn't have a headache that made them feel terrible, but once again the headaches, the distribution is different, ones related to defecation.

I think it is hard to combine pain although people use globals in a variety of ways, but they can confuse things by including both measures of effectiveness and measures of harm, and they may not be telling you how an individual patient feels about the distribution of those symptoms in that person, which is obviously of interest, but tricky to do.

DR. HIATT: John.

DR. FLACK: I just wanted to follow up on

Dr. Harrington's question, and that is, were the people who were assessing the side effects and the headaches the same as the people who were making, transcribing the pain assessments? Those were the same people doing that in the clinic?

DR. AZARNOFF: The people doing the pain assessments were the subjects.

DR. FLACK: Okay, but then the information was transcribed, passed on to someone to--

DR. AZARNOFF: Correct. The subjects made a hash mark. The distance from the left end to the hash mark was measured. That number was transferred to the case report form. Those measurements were checked randomly, they were audited in addition, and in some instances, a 100 percent check was done.

DR. FLACK: But the same people making that, helping the subjects with that, were the ones determining the headaches?

DR. AZARNOFF: The headaches were evaluated again by the subjects. They recorded whether they had one or not, and they recorded the

severity, not an outside person.

DR. FLACK: Okay.

DR. HIATT: All right. So, I think our purpose here was to clarify some of the information in the clinical data presented, and then coming up will be your statistical analyses and risk/benefit.

I just want to ask, put it to you all on the committee, is anyone interested in a break at this moment in time, or do you want to continue on with your presentations?

DR. AZARNOFF: I think it might be appropriate to hear the statistics first and then have a break.

DR. HIATT: Let's carry on with the statistical presentation then.

Statistical Methods and Analyses

DR. GIBBONS: I am Robert Gibbons and I will try and walk you through some of the statistical issues without inducing a headache.

There are a couple of statistical issues that I think are floating on the floor and let me try to deal with those just upfront. Dr. Flack

raised the important point that mixed-effect regression models, generalized mixed-effect regression models are capable of including time-varying covariates to adjust the outcome variable for potential confounds.

In fact, in our sensitivity analyses, we used that benefit of the analysis, and we did the following things.

We looked at headache, any headache, not NTG related headache, but wherever it occurred as a time-varying covariate from day-to-day basis. We looked at the presence or absence of sitz baths, which actually were in greater frequency in the placebo group relative to the active treatment group.

We also looked at the use of analgesics as A time- varying covariate. In no case was there a significant effect of those time-varying covariates - analgesic use, headache occurring at any point in time and at any level of severity, and the presence or absence of sitz baths or fiber on the average pain ratings or the defecation pain ratings, and in

no case was there a treatment by time-varying covariate interaction, indicating that the presence of headache, the presence of fiber, the presence of sitz bath, the presence of an analgesic did not moderate the relationship between treatment and outcome on anal fissure pain.

Another issue that came up--oh, and about this NTG-related headache, the only case where the definition of an NTG-related headache was used, it was never used in the analysis, it was for the basis of deciding whether or not to do what I think is a statistical mistake, a rather egregious one, use an LOCF imputation in the context of a generalized mixed-effect regression model.

Only those subjects that dropped out of the study for an a priori defined NTG-related headache were imputed using LOCF plus a random error component. Again, we objected to this at the time that it was recommended, I continue to object to it strenuously.

So, in all other analyses involving headache, we did not use this definition of an

NTG-related headache.

Finally, the other issue was the dose response, and you had done some analysis on the initial study looking at overall dose including the effects of frequency and initial dosage.

One of the things that Dr. Hung looked at in his re-analysis of the first study, and we looked at, as well, we pooled the data for the two frequencies, the b.i.d. and the t.i.d. in our original analysis of the prespecified secondary endpoint, which was pain.

We did that because there was no significant frequency by dosage, by time interaction in that study. Dr. Hung noticed that there were differences between the t.i.d. and the b.i.d., and, in fact, in our analysis, we also noticed that there were some differences, most notably that in the t.i.d. condition, where the placebo patients were rubbing cream on themselves three times a day, they seemed to have an improved response, again not statistically significantly so, but an improved response relative to the b.i.d.,

indicating a smaller effect size in the t.i.d. group than we saw in the b.i.d. group.

It was for this reason that Studies 2 and 3 were designed for the b.i.d. frequency. So, it doesn't surprise me that when you go ahead and compute total dosage based on the combination of b.i.d. and t.i.d., you don't see a clear dose-response relationship, because the t.i.d. had a pretty pronounced effect as both we and FDA found, or the placebo group, minimizing the effect size there.

If you take the b.i.d., an apples to apples comparison using the data from both the first study and the second study, that both had multiple doses, placebo 0.2 and 0.4, you see a very clear dose-response relationship which is statistically significant, that is, the dosage by time interaction was statistically significant, and that was the slide that Dr. Azarnoff presented.

So, that is sort of how light can travel as a wave and a particle in the analysis of these data.

DR. HIATT: Is it possible, then, just because you just said a lot just now, just to clarify what you have said, so if you believe that the response is not linear over time, the quadratic term, how do you impute missing data then?

DR. GIBBONS: Well, first of all, I didn't even talk about the quadratic thing yet. I am looking forward to talking about that. But let's get to that.

DR. HIATT: Maybe we should just let you go with your presentation then.

DR. GIBBONS: I mean if you want to talk about the other stuff I rambled on about--it will come up in the context of this, as well.

[Slide.]

We have already talked a little bit about mixed-effect regression models, and in 1998, when we undertook this initially, we had specified a mixed-effect regression model as the method of choice for the analysis of the secondary endpoint.

This was still a little bit early on for the use of these kinds of methods particularly by

FDA. In fact, most of my colleagues in industry, when I would recommend using a mixed-effect regression model, making use of all of the available data, would go, no, no, no, FDA will never buy off on that, you have to use an LOCF analysis.

I said, well, no, FDA has lots of very good statisticians, they understand this stuff, this is what we are going to propose, and, in fact, it was accepted.

We used a mixed-effect regression model for all of the analyses for pain, all prespecified Studies 1, 2, and 3. The advantage of the mixed-effect regression model or a full longitudinal analysis is that it uses all of the available data from all subjects, so that we get a much clearer picture. There is no potential bias by eliminating all of the intermediate time points that the investigators went to the trouble and expense of collecting.

It doesn't rely on a single measurement, the last available measurement to characterize a

subject's response to treatment, and subjects that leave the trial early are not artificially considered to have completed the trial.

We don't assume that somebody who left in the first week of the study would have continued at exactly that same level with no variability whatsoever, and we don't weight that subject as heavily as a subject who actually made it through the entire trial.

So, from a logical perspective, mixed-effect regression models have much to offer particularly in the cases where we have missing data due to dropout.

From a more technical perspective, mixed-effect regression models are unbiased under Missing Completely at Random assumptions and Missing at Random assumptions. Now, that is a bit of a mouthful, but what does that mean?

Well, Missing at Random means that the missing data, those data after dropout are predictable from the available data, that is, both the covariates that are in the model, the fixed

effects and potentially the random effects, as well as the available measurements on the outcome variable while that subject participated in the study.

Now, Missing Completely at Random only conditions on the covariates, so it doesn't let you use the available data to draw an inference to what would happen after the subject dropped out.

That is why people don't use--I mean when GEE came out, brilliant idea, but it operates for the analysis of longitudinal data under the Missing Completely at Random assumption, and mixed-effect regression models, which add in some additional assumptions about distributions, give you the additional ability to condition on the available data, so you can predict what would have happened from what happened to a subject.

So, if a subject is doing really poorly in the study, and then drops out, we would expect that they would continue to do poorly. If a subject is doing very well and drops out, we would expect that they would continue to do well and drop out.

Now, there are other missing data mechanisms between MCAR and MAR. You can also have missing not at random. There could be something else like headache, which could change what would happen to a subject after they dropped out. If you drop out due to a headache, maybe something very different is happening.

One of the nice things about mixed-effect regression models is that we can now test to see whether or not we can do sensitivity analyses using models that are more general, that allow for missing not at random.

Now, potentially, there are thousands of different missing not at random models. You know, it could be alien abduction as one of the potential confounders, anything could happen.

So, there isn't an ultimate insulation from bias due to dropouts that will work in all cases, but we have done this kind of sensitivity analysis specifically looking at headache in mind, and our results indicate that headache is not biasing the results and that Missing at Random is a

very reasonable assumption for these data, and I will show you evidence of that.

[Slide.]

Historically, when people didn't do an endpoint analysis, they did some kind of split plot analysis or what has been coined a "repeated measures" analysis of variance. I think that term comes with the psychological literature.

That analysis operates under some pretty restrictive assumptions with respect to the correlational structure. Statistically, that is called "compound symmetry," and it basically says that the correlation between time points and the variances between time points are all equal.

It is a pretty unrealistic assumption. Time points that are close together are going to be more highly correlated than time points that are further apart.

Also, based on traditional least squares kinds of estimation, the accommodation of missing data and dropouts is quite limited in the traditional least square estimated mixed model or

repeated measures analysis of variances it is also called.

[Slide.]

So, there are several useful things about the way missing data are handled in these models. One, there is no restriction on the number of observations per individual, it's a regression model. So, you know, if I have 34 measurements for one subject and 7 measurements for another, it is not going to care.

The subjects don't even have to be measured at the same points in time. They could be measured on a continuous time scale and that's not an issue at all. We don't need to have a balanced design.

They are not excluded if missing data from a prescribed observation period is not available. We use all the available data. There is absolutely no need to impute missing data. In fact, by imputing missing data, using LOCF and doing this kind of analysis, you lose the benefit of unbiasness under Missing at Random. In fact, you

even lose it under Missing Completely at Random. So, it's a really, really bad thing to do.

Again, as I have told you, the assumption of the model is that the data available for given subjects, both the outcomes and the covariates, are representative of that subject's responses after dropout.

[Slide.]

Let's start to get into the data, because, you know, a lot of good points have been raised about this. These are the data from Study 1. These are the b.i.d. data from Study 1. The first thing that you see is that there is a very clear differentiation between active treatment and placebo. The effects are relatively linear over time, and more importantly, the difference between placebo and treatment is linear over time.

When we first saw this, the primary endpoint was healing and we saw that, boy, there is no difference, and I had to tell my friend Dan that things aren't working out too well. We took a look at the secondary endpoint and saw that there was

pretty good evidence that these patients were in a lot less pain. It was obviously, clinically significant and certainly statistically significant.

If we go back and now by the third study, we are looking at day 21 and reanalyze these data from day 21, we see that there is statistically significant evidence of a treatment-related effect at both days 56 and at days 21.

[Slide.]

Now, this is what happened in the second study. Now, in the second study, we see again through day 21, there is a clear linear trend and that those two linear trends are well differentiated between the treated and controlled subjects. That difference through day 21 is statistically significant. This is the second study which was designed to only have the b.i.d. data, so this is just a b.i.d. study.

It is also statistically significant through day 56, but not in a model that only had a linear term. And why is that? Well, first of all,

there is a little curvilinearity in both groups from day 21 through maybe day, oh, 45 to 50. That is not the issue.

The issue is at the very end of the study, in the last six days of the study, there is no treatment effect, it goes away, so it's not that there is curvilinearity in the response process, the treatment by linear time interaction would still have significant without the quadratic term.

What happens is when you try to draw lines through the differences between the treatment, the delta, difference between active treatment and control, those lines get very close together, because you are trying to hit that target at the end of the study.

So, we put in, we had to put in, it wasn't significant if you didn't put it in, a quadratic term in that model, and a quadratic treatment by time interaction, but we interpreted the same treatment by linear time interaction alone.

Now, statistical guidance in this area would tell you that when you have both linear and

quadratic effects on the model, you really should do an omnibus test and get a likelihood ratio chi-square statistic to tell you whether the inclusion of both of those terms is significant or not.

Well, clearly, that would be even more significant than just relying on the treatment by linear term. Now, Dr. Hung, in his reanalysis of these data, raised the important point, and it's an obvious point, that if you have a higher order polynomial term in the model, if you have a quadratic term in the model, the interpretation of the treatment by linear time interaction is more complicated.

It says that it's not the same over time. That's what the quadratic term does, that's your only interpretation. Now, that can mean a lot of different things. It could mean that on day 1, it is linear and the rest of the time there is nothing there, or it could mean, as you look at the data, that through day 21, it is very linear, and it is still pretty linear between day 21 and day 50, and

that it all goes to hell at day 50 through day 56, and that is what is going on in these data, and that is the mystery of the quadratic effect, and it is really not that complicated.

It also led to our design of the next study using 21 days, because that was the linear portion of the curve that we saw both from Study 1 and from Study 2.

[Slide.]

Now, with respect to Study 3, I am going to start out with the more complicated model. I am going to start out with this issue of, well, you know, it's headache, and it's dropout for headache, and you are assuming Missing at Random, and maybe it's not Missing at Random because the headaches are biasing things.

[Slide.]

So, one of the things we can do now is we can fit a model based on Missing Not at Random, and we can fit the kind of model, which is called a "shared parameter" model. Following the brilliant work of my colleague Hedeker and me, in our new

book "Longitudinal Data Analysis" that is no longer in press, it's actually out on the streets, and you can put it on your coffee table and it's really good at parties, we describe how you can fit this kind of model.

The basic idea of this model is that you are simultaneously fitting a survival model, a time to dropout model, with its own set of covariates that include a subset of the covariates from the outcome model, specifically, the random effects that tell us about intercept and slopes of the individual subjects.

So, those random effects, the rate of change and the baseline score from the outcomes model, from the mixed-effect regression model, are shared with the survival model that predicts time to drop out, and we allow them to interact with headache.

So, headache is in there as a main effect, but it is also in there as an interaction with where people start out, the intercept, and the rate of change over time in VAS scores, average VAS

scores.

So, all of that is simultaneously fit using a program called SAS NLMIXED, and if you think that you are going to get a headache from nitroglycerin, try using SAS NLMIXED. That's a real nightmare.

[Slide.]

The results of the analysis showed that none of the terms related to headache were significantly related to dropout. There is headaches all over the place. Treatment by time interaction was still significant.

Treatment by linear time interaction through 21 days in the shared parameter model that adjusted specifically for the effects of headache and the effects of headache on rate of change was not only still significant, it was even the probability value was lower than for the model that did not include this.

It is certainly not biasing the results, and if anything, it is biasing the results more towards the null than towards the alternative, and

headache is not related to dropout in treatment efficacy.

[Slide.]

The second issue, which I think is also a very good point raised by FDA, is that if there is an increased incidence in headache in the active treatment arm, there will be perhaps an increased likelihood of taking an analgesic, and maybe the effect that we are seeing is due to the analgesic, and not to NTG.

It is a perfectly reasonable alternative hypothesis, so to look at that, first of all, we treated analgesic as a time-varying covariate in that same model, and found no evidence of that, but to try to look at it in an even more sensitive way, or it's really a less sensitive way, but a way that we can communicate it better, we looked at those NTG subjects who had a headache and who didn't have a headache.

So, the ones who took analgesics are the blue line, and the ones who didn't take analgesics are the red line, and as you can see, the ones who

took analgesics actually had worse pain, worse average pain, not the opposite.

[Slide.]

This goes completely in the opposite direction of that hypothesis, and if we now just look at the placebo patients, we see an even bigger effect. Actually, taking an analgesic is probably a sign of having worse overall pain, and, in fact, the patients who took medication had worse average anal fissure pain than those that did not.

Then, if we just condition on the analgesic users, we see that, in fact, when we compare the blue line, which is NTG, versus the red line, which is placebo, we see a very sizable effect, a very large effect. It is something on the order of 15 millimeters on the VAS scale, which is a very large effect given that we are now at scores of around 30 millimeters at the end of the 21 days of this study.

[Slide.]

So, clearly, analgesic use is not confounding this effect. It has nothing to do--we

are not having the situation where people are getting a headache, they take an analgesic, and they are doing better because they are taking the analgesic, in fact, the people who are taking analgesics are in more pain related to their fissure, and they do better on NTG relative to placebo.

[Slide.]

Now, let's get a little bit more into the FDA requirements for Study 3. This is an SPA, it was all negotiated. We had a meeting. We talked about this and basically agreed that we would do the following things.

One was that FDA wanted us to combine sites with fewer than six subjects, and there are lots of ways of doing this, and, of course, it's a good idea. We certainly don't want to have a confound between center and treatment.

Despite our objections, however, they required that we do this last observation carried forward business for those subjects that dropped out for an NTG-induced headache, and we have talked

about this already.

Despite that fact, even when we used last observation carried forward in the mixed-effect regression model, we still get a significant treatment-related effect at the 0.05 level.

If we do what is, in my opinion, a more appropriate analysis using all available data without any kind of imputation, since we don't need to do the imputation in the first place and all it can do is hurt us by making our inferences no longer general under MCAR and MAR, we get 0.03 for the effect, a result that indicates that the difference between treated and control subjects is inconsistent with chance expectations.

This is not a measure of effect size. 0.03 is not, you know, 10 times bigger than 0.003. It allows us to make the binary decision that the results that we have seen are inconsistent with chance expectation, period.

[Slide.]

This gives you an idea of something more about the effect size. What we see is that the

difference between placebo and treatment maximizes at about 15 days. It is about a difference of about 25 percent in the magnitude of those scores, and then it goes down to about a difference of about 18 percent, 15 to 18 percent in the last three days of the study, 18 through 21.

Although not the primary endpoint, we looked at 3-day windows of individual point in time contrasts. We found statistically significant differences between placebo and active treatment between days 13 and 15, days 16 through 18, days 19 through 21, and results that approach statistical significance as early as day 7.

Again, this analysis uses all available data from each subject and is unbiased under MAR.

[Slide.]

In terms of secondary endpoints, we see that through day 56, now with a prespecified quadratic term in the mixed-effect regression model for the long-term analysis, is again statistically significant - defecation pain at 21 days, this more serious pain is significant, and defecation pain at

56 days is statistically significant. Again, consistent evidence from study to study to study that the results that we are seeing are inconsistent with chance expectations.

[Slide.]

If we define a more perhaps clinically interpretable endpoint of time to reaching 50 percent improvement, we see again that there is a numerical superiority. This difference in the time to 50 percent improvement was not statistically significant, but we are seeing a difference of as much as 7 days in terms of pain relief down at the 75th percentile or the 25th, where it is listed here as 0.25 in the survival distributions.

[Slide.]

If we go back to Study 1, we see a huge effect. The time at which it takes 75 percent of the people to reach 50 percent improvement is approximately 20 days on treatment, but never on placebo. We never get to the point where 75 percent of the placebo patients reach 50 percent improvement. That result is statistically

significant.

[Slide.]

If we look at Study 2, we get a very similar kind of effect again, which approaches statistical significance, again a consistent difference whether you look at it in terms of the raw numbers or you look at it in terms of a criterion like 50 percent improvement.

[Slide.]

If we go back and then reanalyze Studies 1 and 2 using exactly the same method and time window, 21 days for the primary endpoint efficacy analysis, that we did in Study 3, we see that it is significant in Study 1, it is significant in Study 2, and it is significant in Study 3, 21 days. This isn't about b.i.d. and t.i.d. This is all b.i.d. This isn't about quadratic terms. This is all a simple linear model evaluating the treatment by linear time interaction through 21 days.

If we go back and look at all 56 days, again, statistically significant in Study 1, in Study 2, and Study 3, and the combined analyses.

Now, a very reasonable combined analysis, we are combining a prespecified secondary endpoint from Study 1, the primary endpoint in Study 2, and the primary endpoint in Study 3, all showing very consistent results.

[Slide.]

If we look at those subjects who had a sentinel pile, in Studies 2 and 3 where this was recorded, again, we see a statistically significant effect on time to 50 percent improvement, and that there is a lot more days that subjects are having a 50 percent improvement on active treatment relative to placebo, they are doing better.

[Slide.]

If we combine all three studies, we see again a very similar picture that is also statistically significant.

[Slide.]

If we now restrict attention to a subgroup, that is, those subjects in moderate to severe pain initially, which we have defined as greater than 50 millimeters, we see that not

only--now, there is 144 subjects out of all three studies who met this criteria--only 17 from Study 1, so there is not a lot of power to detect anything, 35 subjects from Study 2, and 92 subjects from Study 3.

We see overall in this more severe subgroup, there is a statistically significant effect combining all 144 subjects. It maxes out at day 15. The magnitude of the effect is not 3 mm, it is 13.5 mm, a sizable difference.

At day 21, it is still 10 mm. If we break it down by study, it's as high as 22 mm at day 15 for Study 1. If we look at defecation pain, overall it is statistically significant, it's a difference of 6 mm at day 15, 8 mm at day 21, and as large as 26 mm in Study 1.

[Slide.]

If we look at the raw data over time for the two groups without any statistical analysis, this is combining the studies, all 144 subjects, we see that by day 21, there is a difference of about 10 mm. You can see that the lines start to

separate as early as 4 days. Again, every single curve that I have shown you shows the same general pattern time and time again.

[Slide.]

If we break the data down into quintiles, and the advantage of breaking the data down into quintiles is that we preserve a balance of the sample size in each of the five subclassifications, we see that the major effect is in the fourth and the fifth quintiles in the more severely ill patients, and that the effect in the fourth quintile, the little table on the right, is as much as 46 percent for Quintile 4 on day 15, and 45 percent on day 21, and for defecation pain it's a 33 percent reduction in pain, and a 39 percent reduction in pain for day 21.

[Slide.]

If we do the same analysis just for Study 3, we see that in the fourth quintile, these more severely ill patients, 67 percent reduction in pain, 65 percent at day 21, 42 percent at day 15 for defecation pain, and 31 percent at day 21 for

defecation pain. Very large effects. They are somewhat smaller in the fifth quintile when you start to get to the really extreme cases.

[Slide.]

Doing it as time to reach 50 percent improvement, we see that in the active treatment, 75 percent of them reach 50 percent improvement by day 7, whereas, it's a little hard to read, but at best you are at day 21 for the placebo patients. You know, that is a difference of 14 days in terms of patients reporting, self-reporting 50 percent improvement in their symptoms.

[Slide.]

We see a very similar thing when we combine all three of the studies.

[Slide.]

Summary and Conclusions. Acceleration in the rate of change in 24-hour pain intensity over the first 21 days of treatment is significantly better than placebo with or without last observation carried forward imputation using a mixed-effect regression model. The appropriate

analysis is without LOCF imputation. There is not a single reference in the statistical literature that would tell you to do that.

Reanalysis of Studies 1 and 2 for 21-day endpoints were similarly significant.

The major effect of Cellegesic nitroglycerin ointment is in those subjects with moderate to severe anal fissure pain, baseline scores, VAS scores of greater than 50 mm.

Analysis of data from Study 3 provides evidence that headache, dropouts and acetaminophen usage does not affect the efficacy results, further providing validation of the missing at random assumption for the mixed-effect regression model.

I would like to now introduce Jonathan Lund, but I have a feeling you are not going to have him up here for a little while.

DR. HIATT: Maybe we could ask some questions from your presentation first. John.

DR. FLACK: Thank you for that clear presentation. I have a question about the use of analgesics, and you particularly focused on

acetaminophen.

How did you actually put that in the model, was it just you used it or you didn't use it, or did you try to create therapeutic intensity scores or something quantitative?

DR. GIBBONS: My memory of that analysis is what we did is we had a binary indicator of analgesic use on a daily basis. So, you are actually getting a lot of quantitative information in the sense that every single day there was a binary indicator of analgesic use, so it could really range from anything from zero over the course of the study to 21, but at each day it was a binary indicator.

Dan, is that your memory, as well?

DR. AZARNOFF: Yes.

DR. GIBBONS: I have done a few things since that analysis.

DR. FLACK: Secondly, I know in the hypertension world where I am, we tell people all the time to don't take this or don't do that, and quite honestly, they are free living and they can

take as much analgesic, Motrin, acetaminophen as they want, and just because you specify it in the protocol doesn't mean they won't do it, or another physician won't prescribe it.

How confident are you in this analysis that you really are not dealing with a problem of, one, inadequate quantification of drug use, because if you don't really quantify it well enough, you can't control it, and when we work with time-dependent covariates like using therapeutic intensity scores or some binary variable, the TIS score always gives us more information than just a simple yes/no.

DR. GIBBONS: I think I am going to have to defer to my colleagues who designed and carried out the study there, but the data that we used, I mean we look at it in two ways.

One, we looked at it as time-varying covariate, so at least we have the transitions at every single day through 21 days, and we also did it through 56 days, so we have got a lot of movement back and forth, so if there was an effect,

you would expect that at least it would be significant, it might not moderate the treatment by time interaction, but in this case, it was neither.

The second analysis where we actually looked at those people who didn't take any analgesic during the whole course versus those who had taken analgesics, so we kind of looked at it in two different ways that should deal with much of that.

Dan, do you have anything to add to that?

DR. AZARNOFF: No.

DR. HIATT: We will go around the room, too, but in your intro, you talked about dose response. In Study 1, there was a b.i.d. and a t.i.d. dosing, and i think what you said is that you were able to combine the b.i.d. and t.i.d. doses because you couldn't find a difference there.

But you do think that there is a dose response when you go from 0.2 percent to 0.4 percent, so I am trying to understand what that means. In other words, the total daily dose doesn't seem to be the issue, but when you give it

b.i.d., the concentration matters, is that correct?

DR. GIBBONS: That's correct. In terms of the effect size, we didn't really see any difference in the active groups, the t.i.d. and the b.i.d. There was no frequency by dosage, by time interaction, so we didn't see any evidence that doing it three times a day made any difference than doing it two times a day statistically.

When you look at the data in terms of effect sizes, without layering statistical significance on it, what you see is that the placebo patients did a little better on t.i.d.

DR. HIATT: Which suggests if you put the goop on three times a day, just the vehicle, there may be some symptomatic relief. I wanted to ask that question. If you just applied a vehicle every hour, would there be a dose response there?

DR. GIBBONS: I don't know the answer to that, but from what little data we have, and remember the t.i.d. was only done in the first study, it would be a reasonable hypothesis to say the more goop you put on, the less overall average

pain you are going to have. I don't know, we would have to do another study.

DR. HIATT: That was for placebo, is that right?

DR. GIBBONS: That was for placebo. That was the inference that we drew from the first study, and that was what led us to select the b.i.d. dosage, particularly since we didn't find evidence of any improvement with t.i.d. for the active treatment, that's why we selected the b.i.d. dosage, and, in fact, that is what has been done in the second and the third studies.

DR. HIATT: I just think in general, what we try to establish, I think, is some level of dose response for most drug approval, and I think the most information comes from Study 1.

You focused on a b.i.d. dosing and a single dose for all the primary endpoint efficacy data, and I think that is supportable if you have the background information that leads us to believe that higher doses are not tolerated and that lower doses are ineffective, and you have defined the

dose-response range.

But I am still bothered by the total daily dose not being related to the benefit, and the suggestion that if you apply placebo more frequently, that may confound the results, as well. Am I correct in those assumptions?

DR. GIBBONS: I am trying to find that slide. Here is the slide. This is the slide that I think addresses that issue.

[Slide.]

So, what this is, is a combination of the 0.2 and the 0.4 in placebo from the first two studies under b.i.d. dosing. What we see is a very clear dose response, so we see that when you keep the frequency of administration constant throughout much of the range, I would say from day 11 through day 56, you are finding that placebo is on the bottom, 0.2 is in the middle, and 0.4 is showing the best improvement, and when we do a minimally effective dose in a mixed-effect regression model, we find that it's the 0.4, that the 0.2 did not separate from placebo, and that the 0.4 did

separate from placebo in this analysis.

DR. HIATT: Just to finish that up, in Study 1, you did test a higher dose than 0.4 percent b.i.d., which was giving 0.4 percent t.i.d., and then at least in that one study, you actually lost significance on the linear response over time.

DR. GIBBONS: We never did a post-hoc analysis where we actually where we actually compared 0.4 b.i.d. to 0.4 t.i.d. We only did the overall analysis in which we looked at the overall time by frequency, by treatment interaction. Admittedly, something the study was probably not powered to detect, and we did not find evidence of it.

It was on the basis of that, that we pooled the b.i.d. and the t.i.d., and in part, it was on the basis of Dr. Hung's reanalysis of those data suggesting that, you know, well, it wasn't significant, we confirmed that, but there were some differences that we saw in terms of the effect size, and you have done the same thing in your

reanalysis.

It was on the basis on that, that we selected a single dosing frequency b.i.d. for doing Studies 2 and Studies 3.

DR. HIATT: Why don't we go around the room. David.

DR. DeMETS: I have three questions I will ask now just for clarification. I am sure we will discuss many of these issues throughout the day.

The first question has to do with rate of change as an outcome, but over what period of time. So, am I correct that in Study 1, you were thinking of a 56-day observation period? I mean it is not clearly specified, but at least that seems to be the implication in what I read.

DR. GIBBONS: That's correct. In Study 1, it was 56 days. In Study 2, it was also 56 days.

DR. DeMETS: And then you changed to 21.

DR. GIBBONS: Then, we changed to 21. The reason we changed to 21 is the FDA said, well, what do you want to do, do you want to prespecify in the third study doing, you know, putting in these

quadratic terms, but if you do that, then, we can't really know where the rate of change is. So, we are a little concerned about that, so if you want to have that, you are going to also have to specify an effect size, so that we can say this is a big enough effect.

We said, well, you know what, the data from the first two studies are clearly linear, at least through 21 days, why don't we design the third study to take benefit of that clear linearity through the first 21 days, and then we don't even have to think about this quadratic term, we will buy off on linearity through the first 21 days, and prespecify it for a secondary endpoint.

DR. DeMETS: So, then, the analysis, the combined analysis or the reanalysis of Studies 1 and 2 at 21 days is really a post--well, I don't know if the term post hoc is the right term--but it's after reflection of the data.

DR. GIBBONS: Very much so. It's taking what we learned in Study 3--well, not taking what we learned--it's taking what we did in Study 3 and

determining did it, in fact, if we had done it that way from the start, would it have replicated in Studies 1 and 2.

DR. TEMPLE: David, we also thought that it was obvious that's where any action, if there was any action, was, and that it was probably most important to patients to get better in the first three weeks, and it was sensible to focus on that.

DR. DeMETS: I am not arguing, I am just trying to clarify that.

DR. TEERLINK: No, but that is part of why we suggested at 21 days.

DR. DeMETS: My second question, I am not sure if you or somebody can help me, I am not used to dealing with Visual Analogue Scales much.

DR. GIBBONS: Don't.

DR. DeMETS: I am used to much more definitive outcomes, but if I understand, at one end of the scale, you have zero, no pain, nothing. On the other hand, you have the worst imaginable, and in between there are no landmarks, there are no benchmarks.

So, I am trying to understand what it means to have--we were putting a lot of emphasis on this outcome in the whole discussion, right?

DR. GIBBONS: Yes, sir.

DR. DeMETS: I don't understand the outcome. What does it mean to have 5 millimeters, or whatever, difference in some slash marks? What does that translate into? So, that is one question. Two, given that somebody believes this instrument has meaning, what do we know about it? That scale is not unique to this study, I assume.

DR. GIBBONS: There is a huge history of using this.

DR. DeMETS: Can you help me understand what the validation and history of this instrument?

DR. GIBBONS: I can give you--

DR. DeMETS: A brief one.

DR. GIBBONS: --a brief smattering of what I know about it, and then my colleagues can certainly give you a much more thorough discussion, but let's see if I can give you enough of it.

Yes, it's a very unstructured scale. I

would probably, if someone came to me today wanting to design a study to look at pain, I would probably guide them away from this VAS scale as quickly as possible, but it does give you probably an ordinal form of measurement.

Obviously, one person's 50 is another person's 30, and there is a lot of inter-individual variability. There is intra-individual variability, as well. One of the reasons why we selected a mixed-effect regression model in the first place is the model is based on the idea that these are individual differences in both the intercept and slope of the time trends, so it seemed well suited to something that was a pretty subjective rating.

I think if we were to do it again, and use the VAS scale, we would have given anchor points along the baseline that would have given more the sense of an ordinal measurement, but if we are going to do that, we probably would have been better off just using an ordinal scale with very concrete, you know, this is the worst pain you have

ever had, this is, you know, the pain of listening to a statistics lecture, this is, you know, whatever, and laid it out so it would be better information, and we could have made category kinds of comparisons.

So, the problem is that we saw it in the first study, we used it in the first study as a secondary endpoint. Once we did that, we sort of had to keep it around in order to use the second study as a second pivotal study using the same endpoint, because we really couldn't change to something else.

So, all of the limitations that you are noting about it are certain valid. It has been used widely in many, any areas of medical research. I mean this is not a new idea. It has been around for a very, very long time. There is a lot of history using it.

There is a lot of cross validations. You are going to see that VAS scale scores are, in fact, related to domains of the SF36, very significant covariation correlation with SF36

quality of life kinds of measures.

So, there is a lot of history in using this. My experience with it, I don't like it.

DR. DeMETS: My third question or comment for clarification has to do with your analysis to assess whether the data are confounded with analgesics or other kind of use.

It has been my experience and also what I have written and published and talked about, that while it is common for us to use time-varying covariates in analysis of data in epidemiology, we have nothing else, we have no other choice, we were trying to figure out do changes along the way predict outcome, but my experience has been in clinical trials, that is a rather risky thing to do.

You are adjusting one outcome for another outcome. I mean compliance is an outcome, concomitant medications usage is an outcome. You start adjusting for things, you can get some very strange and weird results without too much difficulty.

So, while it is interesting what you found, for me, I am not sure that I am comforted that we have ruled out possibilities of that kind of confounding.

DR. GIBBONS: Let me just speak to that, and, of course, I agree with what you are saying. First of all, this was not a primary analysis, this was a sensitivity analysis, so we certainly in the primary analyses did not specify any time-varying covariates, but once we found the overall effect, we wanted to know was it moderated by analgesic use, was it potentially moderated by headaches, was it moderated by concomitant therapy, sitz baths, and fiber.

In all of those analyses, not only did we see no effect of the time-varying covariate on the outcome, but we never saw any change in the treatment by time interaction, which I think in some ways is the more important issue.

Secondly, focusing even more on headache, when we fit the shared parameter model, again, we saw no influence of including a dropout model,

basically, a Heckman selection kind of model, for the effect of headache on dropout and on pain, we saw no effect of the treatment by time interaction for average fissure pain.

So, I am pretty comforted that we have done everything we can statistically to rule out confounds with at least these three groups of potential confounders.

DR. DeMETS: I have one more question. In all three studies, I noted that the number of patients analyzed were not the number of patients randomized. There is always more patients in each study that were randomized into the study.

As someone who had something to do with the idea of the intention-to-treat principle, it's a very simple principle. It says account for all the patients that you randomized, period.

I understand missing data, but my question to you is what can you tell us about the impact of having fewer patients in analysis than you had randomized?

DR. GIBBONS: I agree with you in terms of

the principles of intent to treat. Certainly, one of the nice things about the mixed-effect regression model is even a patient who is only available for baseline contributes something to our variance estimate of the intercept, so there is no reason to exclude such a patient.

I can tell you that I excluded no patients in my analysis of the data. I am not exactly sure of the source of what you are talking about. I know that there was a Russian center in Study 3 that there were some problems with that is described.

There were two subjects that were eliminated from there. I would be happy to look further at it, but I know that from the data that I received, I excluded no one even if they only had a baseline measurement.

DR. HIATT: As we are discussing this approach to the data, I wonder if you all from the FDA could help us understand a simple point.

Usually, a drug therapy is proposed to work at the end of a study, and you count the

number of people who have achieved an endpoint at the end on drug versus placebo, and my question to Dr. Stockbridge and Temple is have you used a mixed-effects model for drug approval where we are talking about an endpoint here, it's a rate constant. It's essentially patients are getting better quicker for a short period of time, and that's clinically how you would view this therapy. At the end, we would all agree they all look the same.

My question is, is this an appropriate way to look at therapy for drug approval?

DR. TEMPLE: It's a good question actually. My light was on because I wanted to raise it. What is most common, for example, in, I don't know, a depression trial, is to ask how people compare at four weeks, but, of course, a lot of people drop out along the way, so you have to do something to deal with the people who left early.

LOCF is commonly used. Everybody dumps on it as being stupid, and we all know it is sort of stupid, but it is not clear the alternatives give

you very different results, and that is something that is under active looking at.

We didn't think it was wrong headed to ask whether the rate of improvement, to look at the rate of improvement, and in a symptomatic situation, that doesn't seem any less inherently reasonable than improvement at four weeks.

I gather one of the advantages of it is it has some of the advantages of a repeated measures test, that you are using a lot of data and can probably detect effects that you might miss if you just looked at the four-week value, but that seems neither here nor there either.

The other thing I guess I noticed, also, and David should correct me, it seems to me that in outcome studies, mortality studies, we take an approach that is not so different from that.

You could look at two-year survival in a trial, but we don't, we look at hazard ratio along the way using the data, and not trying to carry last observations forward. That would be fairly silly for survival.

So, it seems at least similar to this. I tried that out on Jim, and he didn't think it was, but it still seems sort of similar to look at hazard ratio, risk reduction, or something like that, which is basically what you are doing here. You are looking at risk reduction.

I guess I also noticed, and I wanted to ask about this, that having established what you think is an effect on the rate of change, you feel at least somewhat free to look at other things like time to 50 percent reduction. You might even feel free to look at values at 21 days and things like that.

I have to tell you that is something we have been toying with. It is sort of what I was saying before, that if you establish your endpoint overall, maybe that gives some license to look at other kinds of things like take a cut at two years, take a cut at one year without being as statistically rigorous as you would insist on if that were your primary endpoint. I would be very interested in hearing what people say about that.

I am also interested in individual responses. If you have to win on every statistical test, it becomes impossible to do those things. If somehow you got permission to look at those things in a descriptive way, that might be something we should be thinking about, and I think we will be thinking about it, but I am curious of what both Dr. Gibbons and Dr. DeMets think about that.

But the short answer is I don't think there is anything wrong with looking at rate of improvement in pain. That seems reasonable, and I also I guess want to agree that these Visual Analogue Scales are extremely common. Whether you get better by putting landmarks on them or not, I don't know. It is not so different from a scale of 1 to 7 on something that has fewer points on it, but I have a feeling people fill them out more or less the same way unless the landmarks are really good.

Whether they are really good pain landmarks, I am not sure. Maybe for this kind of pain there might be, but in general, it is sort of

hard to say. What you want is for it to be linear over the whole range of it. I don't know how you achieve that. It is not even clear what linear means when you are talking about pain.

DR. HIATT: I just would like to follow up on a couple of things. The thing I am most struggling with is, is it okay to use a mixed-effects model to describe that rate of change. I think that is a trend towards a different approach to the data.

I did want to kind of wrestle that one out. The other thing, just to clarify what you are saying, I think we all understand at the table is, is that you would say to a patient, and this is a question that will come up, in the end, you are all going to feel better no matter what, but you will feel a little better a little quicker if you take this drug.

DR. TEMPLE: Yes, in one way or another, the test we use, I mean if you have a treatment for an upper respiratory infection, you know that by 12 days, everybody is going to get better, so you

arbitrarily pick at one day or two days or time to, but it is not uncommon to have time to a certain amount of improvement. That is perfectly reasonable, so I don't see any trouble with measuring the rate of improvement especially in a condition that is improving by itself.

DR. HIATT: The other question is I like mixed-effects models, because they do use all the data, and you can take the data at different time points, you know, it is locked into a certain schedule. It seems robust in that sense. I think once again, just to understand whether that's an approvable kind of approach to the data.

DR. TEMPLE: Well, we try not to be unduly rigid and accept things that work, but some of the criticisms of the naive version, LOCF, involved in one way or another trying to extend the data to the end by making use of the rest of the data. I don't know if that is the same as mixed-effect modeling, but it seems to bear some relation to it. You are trying to use the rest of the data to fill in the missing people, and I guess I like the idea of just

using the data you have and looking at the rate over time. It has some attractiveness to it.

But, you know, Jim Hung may want to say something about it, because I don't really know what I am talking about in detail here. I am just giving a conceptual idea.

DR. GIBBONS: Can I just speak to the question that you raised just briefly about looking at individual points in time? I am totally in agreement with the idea that once you do the overall analysis, and you either are doing an overall test, and certainly we could do mixed-effect regression models with simple contrast between every single time point and baseline.

So, we could have had a 20 degree of freedom treatment by time interaction, and that would be an example of a mixed-effect regression model, which would allow us to test any deviation across those 21 days. That is a very acceptable practice.

I felt that it was better to talk about rate of change over time in a linear way, or I used

one degree of freedom to characterize the difference between the two groups, which I thought was a more specific kind of effect.

Once you pass on that, I think not only is it reasonable, I think you should go back and then do point in time comparisons. I think they should be done in the context of the mixed model, so that is exactly what we did here. We did these three-day windows to say, well, where is the effect.

You know, it might be that we didn't find any significant effects at any individual points in time, but, in fact, we found them at most of the points in time, and they approached significance, and I am a lot happier talking about something approaching significance when it is a post-hoc analysis.

This is a post-hoc analysis. These individual point in time comparisons are post hoc given that we found a significant effect overall, and not only can they be interpreted, they should be interpreted.

DR. TEMPLE: Bill, just to say something I didn't say, but that Dr. Gibbons is hinting at, you don't save a study that didn't work on its primary endpoint by finding a three-day window that works.

DR. DeMETS: I just want to comment that the idea of a mixed-effects model is not new. I mean we have been using this idea for 35--it was the first thing I learned how to do when I came to NIH. Since then, we have filled in a lot of the theoretical surroundings, so it shows up now in the statistical literature, but it's an old idea, and has the appeal for all the reasons you said.

The tricky part, though, it is not the missing data problem in general, it's the withdrawal and the dropout assumptions that Dr. Gibbons addressed, and that is really where the debate takes place, what do you do with that patient who disappeared.

You make an assumption, was it at random, if they were there, would they look just like the projection, or is there something more, the analogy being in survival follow-up, the patient who gets

censored, do you believe that was random censoring, or was it informative censoring.

So, we tend to try to get as much follow-up data as we can, because we don't believe that the censoring is random. The mathematical theory says that's how you would proceed, but we don't believe that, so we insist on getting as much follow-up as we can in the survival state.

Well, here we have the same challenge to think about, the missing data at the end, was it informative or not. So, that is really the challenge, which is why Dr. Gibbons tried to address that, but the model itself, that's not the controversy.

The issue of what length of time is an issue, which obviously we have discussed, but if you can focus it to the linear part, the first three weeks, that's pretty straightforward I think, so for me, the real issue is do you believe that the patients who drop out, it is really random, or is there something really informative about that.

DR. PICKERING: I have two questions. The

first one has to do with the subjective anal pain ratings. The subjects were supposed to fill it out twice per day, one for average pain and one for defecation pain, so there have sort of been I guess 112 ratings per subject.

Can you tell us what the number actually was, and also how much did it change over time? I would guess that they were pretty good on day 1 and day 2, but a lot of the graphs showed a sort of loss of any effect after about day 50. Could that be because the subjects were less compliant with filling out their pain scales, they knew the study was about to come to an end, that sort of thing?

DR. GIBBONS: I am going to refer that question to Dan Azarnoff. Dan, you are going to have to come here and use the mike.

DR. AZARNOFF: I don't have the exact data with us, but we collected somewhere between 70 and 85 percent of all of the daily measurements.

DR. PICKERING: How about changes over time, I mean at the beginning and the end of the study?

DR. GIBBONS: As we have seen in the analyses, in the first study, through day 56, the difference that we were seeing between active treatment and placebo was maintained.

In the second study, it went away by day 56. Now, there were additional concomitant therapies going on that may have accelerated the rate of spontaneous remission or decrease in pain.

The third study, you know, the effect was again largest through--actually, it was largest at day 15, and by 56 days, the effects were generally smaller. I don't believe that it was due to attrition although the graphs could be influenced.

Obviously, these graphs are the available subjects at each one of these points in time, whereas, the analysis uses whatever is available, so those graphs are not imputing anything. It is just whoever who is around at that point in time.

It may be that there is differential dropout between those people who didn't do very well, so what you are looking at the placebo patients who stuck around to day 56, and they are

doing a little better than everyone else.

So, that could have an effect on the very end of it, but it wouldn't have an effect on the statistical analysis of those data, only the graphical display.

DR. PICKERING: So, if it was 70 to 80 percent, there were about 25 percent missing data. With the LOCF, those were imputed, those pain ratings were imputed from the previous day's ratings, is that how it worked?

DR. GIBBONS: It would have been for those subjects that dropped out in Study 3 for an NTG-related headache.

DR. PICKERING: I am not referring to the dropouts, just people who missed out, you know, a couple of days filling out their--

DR. GIBBONS: Are you referring to the graph or the analysis?

DR. PICKERING: The analysis.

DR. GIBBONS: No, the analysis does not imputation whatsoever. The analysis uses all the available data for a subject. A subject who was in

the study for one day will still be in the analysis, but will be obviously down-weighted. They are not going to be used in the model to differentiate treatments, but they will be used in the model to characterize the variability at baseline and to estimate the mean at baseline, but a subject who has, you know, 56 days or 21 days of information, and another subject who has 3 days of information, will not be treated as being equal weights of evidence.

That is really in many ways a critically important distinction between the LOCF. The LOCF basically says no matter--you know, it follows the intent-to-treat principle, which is a good idea, but it says I am not going to differentiate in any way between a patient who gave me a full 21 days of information and a patient who only gave me 3 days of information, and that is a big problem.

The other problem with LOCF is that there has been a contention that it's a conservative procedure, that it will always kind of hug the null, and, in fact, that has been shown not to be

the case. In fact, the bias you can get from LOCF can go in either direction.

DR. PICKERING: Thank you. My second question has to do with the change over time. In just about every curve you showed, there is a quite pronounced reduction in symptom severity in both groups, that is most marked between days 1 and days 2, and there is something about being admitted to the trial that results in improved symptoms whether it is a regression to the main or placebo effect, or I don't know, could it be that there is some effect of massaging the fissure--

DR. GIBBONS: Whatever that thing is.

DR. PICKERING: Right.

DR. GIBBONS: May it never happen to us.

DR. PICKERING: For example, you showed us the quintiles based on baseline severity of pain, and in the ones with the most severe pain, there were huge changes over time in both groups that were much bigger than the changes between the nitroglycerin and the placebo groups, the differences between the nitroglycerin and the

placebo groups, which to me says that the placebo actually did pretty well, particularly in patients with the most severe pain.

DR. GIBBONS: I think there is lots of evidence for that in any kind of cyclical disease. I mean the one that I am most familiar with is the treatment of depression, and there are pronounced placebo effects in the treatment of depression.

People just get better over these trials, and the magnitude of even our very best antidepressant medications pales by comparison to the overall amount of change on placebo. Yet, it is a statistically significant advantage for these patients to be on antidepressant medications.

So, yeah, definitely, people are getting better in terms of their ratings. It can be for a wide variety of reasons including regression toward the mean. It can be because they are massaging something and getting better. It could be the nature of the disease, that it just gets better over time. There could be a selection effect of, you know, the dropouts running off and getting

surgery. There aren't that many dropouts actually overall in the study.

But the key point is all of those things affect both groups, and the question is what is the added benefit of taking the drug, and is that added benefit consistent with chance expectations or not.

DR. HARRINGTON: Dr. Gibbons, you will have to pardon me if my clinical questions here are a little simplistic, but see if you can help me understand the benefits here.

DR. GIBBONS: You should always beware when anyone starts out like that. R.A. Fisher, when he would talk to a student, who was Gossart, worked for the Guiness Brewery, would always say it follows directly that... and Gossart knew that he was in for two weeks' worth of work.

DR. HARRINGTON: Well, hopefully, it won't take you that long to answer my question.

When I listened to your very elegant presentation, I am impressed with the use of the modeling to look at this sort of data, because of the ability, as you rightly say, to incorporate all

of the information that is available. That seems to be an attractive methodology.

But I am also struck by the fact that with all of this data available, that based upon some of the assumptions one way or another, p-values shift from 0.03 to 0.12, so it suggests to me that with minor manipulations and assumption, that the effect is really at the margin of statistical significance. You very carefully in your point about depression, said patients are statistically better to take this medicine.

As a clinician, I am not interested if they are statistically better, I am interested if they are clinically better. So, help me understand, when you used the Visual Analogue Scale to calculate the sample size, is the benefit that you are looking for, as it is driven obviously by the statistical assumptions, it is pretty small given the amount of data that you have available in this model to discern that difference.

Would that be a correct statement? And I am struck by the fact that with all of the--I like

David's term "post-reflection" of the data analyses--did you do any post-reflection of the data analysis of the clinical outcomes? Our colorectal colleague today said that he has observed that nitroglycerin paste takes him from a 70 percent operation rate to a 20 percent. That sounds pretty good.

Do you have any analyses of the clinical outcomes that might help me look through this?

DR. GIBBONS: I wish that we did have those kinds of clinical endpoints, but they were unavailable, and there were HIPAA issues, and doing long-term follow-up of these patients was not available in the third study.

In terms of the effect size, first of all, in designing the studies, empowering the studies, we used the mixed-effect regression model from the previous analyses, so they were designed to detect the same difference in rate of change over the first 21 days.

So, we used that model, we got those estimates, and we did the power computations on the

basis of those estimates, not on any kind of effect size.

DR. HARRINGTON: But tell me, but there is a delta.

DR. GIBBONS: Oh, yes.

DR. HARRINGTON: So, what is the delta on the Visual Analogue Scale that the trial was powered off of?

DR. GIBBONS: To be honest with you, I can't remember, but it was probably something on the order through 21 days of probably a rate that was fairly close to 1. You know, it was probably between a half and 1, so that a unit change per day of a half a unit or 1 unit, something in that general ballpark. I don't remember the specifics. I could find that for you.

I think that when you look at the data, what you see is overall, the effect is not huge, but as you start to look from subject to subject, and you start to look in the more severely impaired subjects, the size of the effect is really quite large.

You know, it is ranging up as much as 40 or 50 or 60 percent difference between treated and control subjects, so I think there really is a sizable benefit particularly in those more severely impair subjects at baseline.

DR. HIATT: Just to clarify that question, though, Bob, I think--and help me here, Dr. Temple and Stockbridge--generally, the question on the table for these meetings is does drug beat placebo. It is generally a statistical question. How much it beats placebo and whether that is clinically relevant is generally not something we can address. Am I right about that?

DR. TEMPLE: Well, I am going to talk about that a little bit. You are certainly right, that is the usual question that you are asked, and I have got a few slides on effect size. I mean not to state the obvious or what I am going to state later, if we really cared about effect size, we would have a different null hypothesis, I believe.

So, we sort of care about it, but we don't really act as if we do very well, and beating

nothing, even if it's just the lower bound of the confidence interval, is usually enough, but that doesn't mean that is the only possible conclusion.

For example, just to state the obvious, if the drug has toxicity, then, that might not be enough, and we might impose further burdens on a drug, like being better than the available therapy, or being effective in people that don't respond, but that is really reserved for drugs that are toxic.

One of the questions Norm wanted to raise here was how little is too little, or should we be thinking about that, and there is not a really terrific legal or regulatory history on those matters. I will tell you what there is and what we have typically done, but that is part of what we are asking you, is there something so minute that it really is below the level, which is not the position we usually take.

DR. HIATT: I really appreciate that. I just wanted to clarify that point. I know we are going to come back to it, but it kind of came up at

this point in time, and I appreciate that perspective.

Also, we are getting a little close to maybe actually absolutely needing a break. Would it be all right if we could maybe just pause for a second, and maybe continue with you at the podium, and then we have the risk/benefit, would everyone be okay with that? About 15 minutes.

[Break.]

DR. HIATT: If we could resume with perhaps where we left off, with any ongoing questions.

Dr. Koltun, are you still there?

DR. KOLTUN: I am still here, just hanging on.

DR. HIATT: Wow. Do you have any questions? We should give you the floor for a minute.

DR. KOLTUN: Of the biostatisticians? No.

DR. HIATT: Okay. John.

DR. TEERLINK: There are two things I would like to clarify. One, you had suggested that

there wasn't data on operation rates. I think, at least through the first 56 days, there should be information on how many patients were operated, I believe.

DR. GIBBONS: The question is were there operations during the 56 days. There were no operations during the first 56 days.

DR. TEERLINK: So, in general, even though we have been told that this patient population has a high rate, the placebo group would have a very high operative rate, in a surgical practice, in general, this was a very low operative rate in all of these patients, and were these patients enrolled in surgical centers or through gastroenterologists, or who were the main principal investigators? I know this is harder for you. It is not really a statistical question.

DR. GIBBONS: I will echo it over there.

DR. AZARNOFF: The vast majority of the investigators were colorectal surgeons, and a few of them were gastroenterologists, so they are all knowledgeable about anal fissures.

DR. TEERLINK: Right. So, these were patients, these are similar to the colorectal surgeons patient populations where we heard that usually, in the placebo group, there is a very high rate of surgery intervention, yet, in this trial, we saw zero surgical interventions, which I think is interesting in terms of trying to figure out what the patient population group is.

Secondly, in terms of trying to figure out the--

DR. GIBBONS: I am not sure that that is a function of, you know, that this is a special population or being enrolled in a randomized clinical trial where everybody believes that they have equal likelihood of receiving a treatment that may be efficacious during this period.

They all can go after the trial is over and get surgery. So, there may be a disincentive to pursue surgery early on.

DR. TEERLINK: Yes, and that is a possible explanation although if we hear that usually, the surgery occurs within the first 2 weeks, 4 weeks,

waiting until 56 days or 8 weeks is considerably beyond what we would see, at least what our clinical colleagues, surgical colleagues told us.

In terms of trying to figure out, and I guess we will get to this a bit in the risk/benefit, when you were powering the trial, you clearly powered it to be, tried to design it to be a positive trial, and there must have been some consideration, though, in terms of what you felt the minimally clinically significant difference was going to be.

At least in terms of Visual Analogue Scales for other areas, such as dyspnea, which is the area that I am most familiar with, we actually had a sense of what we felt was a minimally clinically significant difference, so a change of 22 mm on a VAS dyspnea score correlated with a moderate to markedly improved dyspnea when patients compared a Likert to another.

So, usually, that is a process that people have to go through to validate an endpoint before they submit for regulatory submission, to say this

is a useful endpoint.

How has this VAS been validated for rectal pain, and what does a 3 mm or a 22 mm difference correlate to clinically? Do we have a sense of that?

DR. GIBBONS: Actually, you are asking two different questions, but the first question is how did I power the study and did I use an effect size, and if I didn't, why not, and the answer is that the initial study was powered with the primary endpoint being healing.

So, we went to the literature, saw this brilliant work from this Lund guy, and noted that he had seen, what was it, like 64 versus 8 percent healing in his sample, and powered the study accordingly, so the study was not powered for the secondary endpoint, which was the pain reduction.

Obviously, we didn't replicate those findings, so when we looked at the effect that we observed, which was significant at some ridiculously low level in terms of probability, we simply used that mixed-effect regression model to

power the second study, and then used that mixed-effect regression model through 21 days to power the third study.

So, the truth of the matter is I never even considered what the effect size would be at any point in time, but only the rates of change that I found to be statistically significant in the previous studies, and powered it to ensure that we would be able to, if nothing else changed, detect that same level, you know, detect a significant effect again using the prespecified methodology.

So, it really wasn't designed to--you know, you can interpret what it would be, I mean if I tell you what the rate of change is, you can figure out what the difference would be at 21 days, because it's a linear model, so that is a direct byproduct, but that was not my thinking in it.

The second question was about correlates of the VAS score, and we are going to see a presentation about correlation, what is the VAS score really measuring, and is it valid, is it related to something else that may be useful for

the purpose of interpretation.

What you will see is a presentation of correlates of the VAS score for anal fissure pain with SF36 domain scores, which are telling us about physical and emotional quality of life, and things like that, and there are all significant correlations with that.

So, it is certainly related to overall well-being, what does it mean for an individual subject is a little unclear.

DR. TEERLINK: Finally, if you could show--you know, we are all careful with subgroup analyses--but if you could show us the subgroup analyses by country in terms of the effects of this, and take your pick in terms of how you want to show it, but for the primary endpoint, how well did this ointment work in the United States compared to Serbia and those different areas.

DR. GIBBONS: We haven't undertaken such an analysis, because in some extent it is confounded with study, as Dr. Azarnoff pointed out.

The first study, which really showed the

largest effect size, was a completely U.S. study. The second study was a mix of U.S. and non-U.S. sites, and the third study again was a mixture of U.S. and non-U.S. sites.

So, we did see--I mean there is no replication, I can't give you a statistical inference, and if I were to do an analysis, it would be confounded by study--but there does appear to be that, you know, the biggest effect we saw was in the U.S. from the first study.

Certainly, we could do an analysis in the second two studies and contrast what we saw in the first, you know, in the U.S. centers versus the non-U.S. centers, but obviously, those would not be powered for that.

DR. TEERLINK: So, are you suggesting that the progressive dilution of effect size that we see through the trials is due to it moving away from the U.S.?

DR. GIBBONS: It is certainly one potential explanation. I wouldn't suggest that that is what is going on. I think there may be a

mixture of things including the use of concomitant therapies, the use of recruitment in centers that were not U.S. centers. Do I think that that is why, I don't know, it is certainly a possibility.

DR. TEERLINK: So, the treatment effect stayed constant at least through the U.S. sites, through the trials.

DR. GIBBONS: If you look at the graphs, the treatment effect is exactly the same both in terms of healing and in terms of VAS pain scores from study to study to study. The only thing that changes is placebo. I mean that is a very, very clear finding. You look at those curves.

DR. TEERLINK: Within the United States.

DR. GIBBONS: Right.

DR. TEERLINK: That is different than the FDA analysis seemed to suggest.

DR. GIBBONS: Well, I think you will see from here that we don't always agree.

DR. TEERLINK: In fact, if you see on page 25, they actually suggest that the only subgroup difference they found was that the substantial

improvement in pain scores with nitroglycerin was in Serbia, and that U.S. patients fared better with placebo.

DR. GIBBONS: Well, you know, that is certainly inconsistent with the data from the first study, which showed the largest effects, that, you know, had no non-U.S. centers, so I don't see what the basis would be for that.

DR. TEERLINK: Just looking at the main pivotal study, Study 3.

DR. GIBBONS: Well, you know, I haven't done those analyses.

DR. HIATT: Related to sample size, at least it is mentioned for the second study, you powered at 80 percent, I wasn't sure how you powered the third study, but one general question is these studies are relatively small, and they don't look terribly hard or expensive to execute.

My question is, if the handling of dropouts really can change the results in Study 3, and if the modeling could potentially affect the results in Study 2, why you wouldn't just overcome

that with a much bigger sample size. I mean why not power it to 95 percent, which might have allayed some of these concerns.

DR. GIBBONS: You power the study based on an effect size that is observed. We never know the true effect size. Then, we pretend in any statistical power analysis that that effect size is, in fact, the true effect size.

It has its own distribution, it is going to vary, and, yeah, we could have said instead, let's try and make up for that by going to 0.8, from 0.8 to 0.95 or let's take double the sample size and really make sure that we nail it.

But based on all of the data that we had, we powered it to a way where we felt that we could confidently make this binary decision of whether or not it was significant, and we were able to do that.

I am sure there were financial constraints that were--you know, this was the third study. None of these people wanted to do a third study. Nineteen countries and Great Britain approved this

without a third study. I think there were limited resources and wanted to do it that way.

DR. HIATT: Just kind of the usual thing obviously, and I guess the consequence will be, you know, how we interpret these, particularly the comment about if handling of dropouts could influence results or the modeling might change it, you know, how clean that signal is.

DR. GIBBONS: Let me speak to that for just a second, because I think that's an important point. Remember that, you say there is one or two data points, well, every subject contributes as much as 56, and for the primary analysis, 21 data points.

Now, if you have a subject who is in the study for a couple of days, and then you either impute or use some post-discontinuation data where the person is off doing God knows what, maybe the person had surgery, and those data are very aberrant from what was observed for the first three days, then, that could have a large effect on the residual variance.

So, it can add quite a bit of noise. It is not like just putting one dot on a graph. You are putting 17, 18, 19 points on that graph, and it is going to change the variance/covariance structure of these random effect models. They are going to be somewhat sensitive.

These models are, in fact, conservative. They regress weird results back to the mean, but you still can have an influence by putting in some values that are very aberrant, and I think that is one example of what has happened.

DR. KOLTUN: This is Dr. Koltun on the phone. Can I change my mind and ask a question?

DR. GIBBONS: Sorry, no.

[Laughter.]

DR. KOLTUN: I am not a biostatistician, but I am a clinician, and the way I interpret these studies and the conclusions that you are explaining are that the placebo groups versus the study groups did not have an absolute statistically significant improvement in the treatment groups for the primary endpoint of pain, but rather it was the rate of

change or rate of improvement of that pain, correct?

DR. GIBBONS: The only analyses that were done and prespecified were in terms of rate of improvement in pain over time. Those were the primary endpoints. No place was there a specification of absolute pain at day 21 should be, you know, different more than 10 millimeters or something like that.

We did, however, once we did the overall analysis and looked at rate of change over time differentially in the treated versus control subjects, do post-hoc, point in time comparisons in Study 3. We actually did them in the other studies, as well, where we took a 3-day moving window and made comparisons at each one of those points in time.

If memory serves me, they were statistically significant in and of themselves starting in day 13 through day 21, and approaching significance meaning between 0.05 and 0.10 as early as day 7.

DR. KOLTUN: I thought the first study, though, the Study No. 1 had no difference in absolute pain between the two groups, but it was the rate of pain that was different, and then that was used as the basis for the primary endpoint target for the subsequent studies.

DR. GIBBONS: That is actually incorrect, as well. The primary analysis for the secondary endpoint, which was pain in the first study, was again a mixed-effect regression model that is comparing the rate of change between active and control subjects over the entire course of the study, but we also did, in that study, post-hoc comparisons to identify at what points in time were there statistically significant differences in the absolute levels of the VAS pain scores, and there were several points throughout the course of that study that were statistically significant, I think beginning very early in the trial, if memory serves me, three or four or five days.

DR. KOLTUN: Oh, really, okay.

DR. GIBBONS: Are you still there? I

guess I answered your question.

DR. HIATT: Do we have other questions?

DR. LINCOFF: Two points. First, related to the follow-up in patients who dropped out, I actually think that your arguments are reasonably persuasive that maybe the LOCF for a treatment such as this may not be the best especially since we are talking about teasing out whether there is actually a signal of a treatment effect, because we are dealing with a symptomatic endpoint that what happens after discontinuation may not be as important.

My question is, in your analysis where you showed the alternative, which was your preferred, that is, without the LOCF, it seems like there were two patients who had post-discontinuation data.

Did you include those patients, because presumably, off therapy, they would be regressing toward the mean, or is this an analysis that only includes patients on therapy? That is my first of two questions.

DR. GIBBONS: The analysis that we

presented that had the probability value of 0.03 for the 21 days used no imputation and no post-discontinuation data, just data, all available data, every single daily point for all subjects while they were taking part in the study.

DR. LINCOFF: The second is, I want to get back to a point that was brought up earlier, I am a little concerned by the fact that as you move from Study 1 to Study 2 to Study 3, as the sample size increases, the apparent treatment effect gets smaller.

Now, I know you said that in the United States that evens out, but I am not sure that is true. I mean I am not sure we are talking about the same thing, because the FDA document says that actually the treatment effect in the United States didn't appear to be as large as elsewhere, so if we just look at the whole studies, unless you actually have U.S.A. data, it does look like the magnitude of treatment effect diminishes.

This, despite the fact that you designed the third study to actually include the higher pain

patients, if I recall correctly, by requiring that they have a score of greater than 50--

DR. GIBBONS: It was greater than 30.

DR. LINCOFF: I am sorry, okay, greater than 30. So, at one point, you did show in your subgroup analysis that the higher pain patients seemed to have more of a treatment effect, and yet, for the overall Study 3, as compared to Studies 2 and 1, there seems to be less treatment effect.

Now, I understand the background therapy, but in Study 2, you actually mandated fairly vigorous background therapy in the control arm where you didn't in this, so I just wonder if you have explanations or if this concerns you that as you get more sample size, as you get a better estimate of the true treatment effect, that treatment effect seems to diminish.

DR. GIBBONS: I think that it may be an over-interpretation to sort of look at diminishing treatment effects and try to figure out what is going on. My view of this is that there are three independent studies. When you look at 21 days of

treatment, you use a model with a linear trend. You get statistically significant differences in every one of them using all the available data.

The effect sizes at any particular point in time are going to vary from study to study, in part due to the composition of that study, in part due how severe in some subgroup analyses, and things like that, which I think are helpful in trying to understand the overall trends, but the critical question is, is there something going on associated with this drug that is reproducible and beyond chance expectations, and I think these analyses are definitive in answering that.

Now, are we seeing exactly the same effect from study to study? No, they are not exactly the same effects. Are the effects in Study 3 smaller than some of the other studies? Overall, yes, they are smaller. They are actually bigger in Study 3 than the other studies in the more severely ill patients.

So, if you just take a look at Study 3 alone, and look at those people who had a baseline

score greater than 50 millimeters, that effect is going to be larger. That quintile analysis is larger for Study 3 than it is for Studies 1 and 2, or the aggregate.

So, there is a lot of things going on and, of course, effect size is going to be a random variable. The critical question is can we interpret something about the effect of the drug consistently over these three trials using the same statistical methodology, using the same frequency of dosing, using the same window of time, and the answer is yes.

DR. WARNER-STEVENSON: As we look towards how we might explain benefit to a patient, I am very attracted to your secondary endpoint of the time to 50 percent improvement in the patients, and then looking at your Slides 74 through 76, I am disturbed by the same trend that we see in primary endpoints, where the curves are very wide for Study 1, narrower for Study 2, and basically, much narrower for Study 3.

As you indicate, that seems to be due to a

better outcome in the placebo group, but I would suggest that it may be that as you went through the trials and had to be more and more precise about documenting the symptoms and improvement in both groups, that it became clearer, in fact, of how frequently improvement occurs in everybody.

DR. GIBBONS: I am not sure there has been any change in the methodology of documentation of symptoms or carrying out these studies. They were all double-blind, they were all, you know, run by similar kinds of monitoring organizations, I don't think there is anything in the methodology at which the studies were carried out that would explain those differences.

They obviously are tapping different populations, there may be selection effects based on the rampant off-label use of this drug. Maybe the patients who you are actually getting into the trials as time goes by and more and more people are being treated by their GI doc with this may change the sampling distribution of where you are finding the patients.

There could be those kinds of things that are accounting for these differences, but I don't think it is anything to do with, you know that we are getting better at the conduct of the study and the results are getting worse, so I would disagree with that point.

DR. HARRINGTON: Are you sure about that, because I thought as I read through the FDA documents, that the way that pain was--let me rephrase that--anal pain was characterized as a secondary endpoint in the first studies, which typically would mean that there may be less rigor applied to it from an operational perspective than if it were a primary endpoint.

Can you tell us that the fullness of the data, which was reported to be 75 to 80 percent complete, is the same over all three studies?

DR. GIBBONS: I believe it was. Dan?

DR. AZARNOFF: Approximately.

DR. GIBBONS: Approximately the same.

DR. HARRINGTON: What does that mean, "approximately"? These are small numbers of

patients, so 60 percent could be a lot different than 65 or 70 percent.

DR. GIBBONS: Do you have that information? We don't have that information with us right now, but we could provide it to you.

Again, these are self-reporting. It is not like the clinicians are being trained better to extirpate the information from these subjects. These are self-reporting. Even though it was a secondary endpoint in the first study, it was using exactly the same scale.

In fact, we went to a lot of trouble to make sure that we were using exactly the same methodology, so that these subsequent studies could be used as pivotal studies, to replicate the results of the earlier studies.

The second study was, in fact, designed as pain as the primary endpoint, as was the third.

DR. DeMETS: I would like to have a question about the withdrawal issue. When the patients had an adverse effect, such as having to withdraw from--I assume that they withdrew from

treatment, did that also mean that they withdrew from the study, or did you have the opportunity to follow these patients after they withdrew from study, and just didn't--

DR. GIBBONS: First of all, the answer is once they were out of the study, they discontinued from the study, but during our design meeting with FDA, they wanted us to collect post-discontinuation data for the purpose of determining whether or not there was a rebound in the anal fissure pain after withdrawal from the study.

That was the sole purpose of obtaining these post-discontinuation data. So, those data were available in a few subjects. I don't remember--about five subjects had some post-discontinuation data, and some of the sensitivity analyses done by FDA, where they indicate that, oh, well, it is not significant, included those post-discontinuation data in their analyses.

DR. DeMETS: Well, the reason I was asking is I would think about this, I would follow rate of

change over, say, 21 days, which is not all that long, that I would have done everything I could to get patient data on all that for that entire length of time, but the problem is we don't have that, which is why we are having this long discussion and computing, guessing, and arguing.

DR. GIBBONS: As a part of the agreement of conducting the study, we did make or they did make every attempt to obtain those post-discontinuation data, again for looking at this very specific question of rebounding effect.

I would be very reluctant, you know, even if all those data were obtained, to necessarily use them in the analysis, maybe in a sensitivity analysis, but to use them in the primary analysis. In fact, they were specified against using those data.

DR. DeMETS: We might debate that point, but later.

DR. GIBBONS: Okay.

DR. PICKERING: I would like to follow up on that, ask you to comment on a statement made in

the FDA statistical analysis. It is actually on page 4, right at the end of the FDA book, but as I understand it, the official primary analysis, prespecified p-value for Study 3 is 0.0498, is that right?

DR. GIBBONS: That's correct. That involves using the LOCF imputation for the subjects who dropped out.

DR. PICKERING: The FDA analysis, this was a subject who was discontinued due to drug-related headache, but did have post-discontinuation data, and according to the FDA analysis, whether or not you include that post-discontinuation data makes it either go from being significant or losing all significance.

If it is just one subject that tips the balance that much, that bothers me.

DR. GIBBONS: Again, you know, it is the same question that you had asked before, it is really an issue of having very little data, and then using data for that particular subject who happened to be very different from the rest of the

responses of other subjects who were similarly treated, and it increased the variance.

I mean it showed there was a lot more variability with the inclusion of that essentially outlier, and it is not surprising it was an outlier, those were data that were after the subject left the study, and I guess it is what it is.

DR. HIATT: As we kind of maybe at some point, we will need to move on to your risk/benefit, but there are some discrepancies--Norm, help me just a minute here--between the FDA interpretation of Studies 2 and 3 and yours.

I think we would all agree Study 1 was a negative trial, right?

DR. GIBBONS: Certainly for the primary endpoint, yes.

DR. HIATT: Yes, which is how you define negative or positive. I think that there are some discrepancies around, you know, whether the--Dr. Hung, for the FDA, evaluating Study 2, found it

negative, and you find it positive, and then Study 3 has this margin of who gets included or not.

So, I think I just would want to ask that before we begin our deliberation, I noticed there is a presentation coming, if we will hear the FDA's perspective on particularly Study 2 and 3 to help clarify this, and you don't have to do it right now.

I mean I don't know when the right time is, but I just wanted to make sure that that got addressed in a more formal way. No?

DR. STOCKBRIDGE: There is no planned FDA presentation other than the little introduction that Bob is going to give. So, if you have got questions, you can ask them now or you can ask them later of the review team that is here.

DR. HIATT: It may be a good time to do that, then, because I think in terms of deliberating, we have to really be able to come from the basis of whether we view these three trials--

DR. STOCKBRIDGE: Before you let Dr.

Gibbons go, I would like to ask one quick question.

DR. GIBBONS: And I would like to start out by maybe, you know, in my own view, what are the differences, what are we saying that is the same and what isn't, and then give Dr. Hung the opportunity to say no, you don't know what you are talking about.

DR. STOCKBRIDGE: Yes, actually, I do think that all of the differences are well understood, so we can have a discussion about where the different p-values come from.

The question I wanted to ask you was in the analysis that you used, from which you concluded that the analgesics did not contribute significantly to the difference in pain scores, can you tell us either in terms of slope or in terms of the effective Visual Analogue Scale at 21 days, what the upper limit of the confidence interval is for the effect of the analgesics, what magnitude of effect of analgesics was ruled out by the analysis that you did?

DR. GIBBONS: Well, of course, I don't

remember, you know, confidence intervals for the third study on this. I mean I don't have that information, but what I do have is that there was no effect, no change, I mean maybe in the third decimal place, on the delta in the slopes between active and placebo when you included analgesic as a time- varying covariate, and then in the subsequent analysis that I presented here, it clearly shows that those patients who were taking analgesic actually were in more overall pain.

The sloped were quite parallel, but were in more pain than those subjects who were not, so that there wasn't an analgesic effect that was decreasing pain whatsoever, in fact, there was a self-selection effect of patients who were in more analgesic--more anal fissure pain taking analgesics.

DR. HIATT: Dr. Hung, I wonder if we could maybe address some questions to you then. So, we have all looked at the FDA version of this, and maybe we could start with Study No. 2. What is your interpretation, is it positive on its primary

endpoint or not?

DR. HUNG: Study No. 2, actually, the sponsor want to focus on the rate of change, and as I said in my review, that the rate of change is not the slope when the model is a quadratic, and that is why, in my view, the Study 2 is not conclusive, but was used informatively to design Study 3.

So, actually, the Study 2 helps to suggest potential treatment benefit for the 7 days, 14 days, 21 days, and that was in my table. So, that was my Study 2.

DR. HIATT: Let's just clarify that. I have to find it, I was looking at your table here, but there is one table I thought you showed that there was a linear trend, not for the 0.2, but for the 0.4, and whether you use a linear or a quadratic equation, you got positive, but then there are other points where I think you have interpreted as negative.

I understand that the quadratic term was not prespecified. I think we all understand that that does help inform us about, you know, because

the quadratic term worked in the model, it allows you to think about the time of the effect and that it may be more dramatic early than late.

I think all that is interpretable, but once again, tell us, your conclusion about Study No. 2 is that it's a negative trial in its primary endpoint?

DR. HUNG: Well, it's not negative, it's suggestive. The data seem to suggest the potential benefit for up to days 21 or 14.

DR. HIATT: Where I am going with this is typically, you want two positive Phase 3's with p less than 0.05.

DR. HUNG: Right.

DR. HIATT: Study 1 fails, so we have to have two positives, 2 and 3, to make it, or one that makes it by a lot. Clearly, 2 and 3 don't make it either of them by a lot.

DR. HUNG: Right.

DR. HIATT: So, we have to then draw for approvability, we have to decide if 2 and 3 are cleanly positive on their prespecified primary

endpoint.

DR. HUNG: Right.

DR. HIATT: And I am confused about that, but I understand when we get to Study 3, it is more about how you impute missing data, but for Study 2, just tell us once more, do you disagree with the sponsor's interpretation that this is a cleanly positive Phase 3?

DR. HUNG: Well, that's true, I disagree. I disagreed because the announced up to day 21, the linear component announces it's post hoc, because the primary analysis was based on the rate of change assuming that the response profile is linear throughout 56 days.

So, based on that, the rate of change is not the primary parameters we are looking at.

DR. HIATT: So, help us understand. Just to clarify for everyone, if you do a complete linear regression across the entire 56 days in Study 2, you are saying that is negative.

DR. HUNG: Of course, there is no statistical significance.

DR. HIATT: And it was only because of the post-hoc inclusion of a quadratic term, it goes from a negative trial to a positive trial.

DR. HUNG: Right, and that helped to suggest that you may be able to see the signal for up to 21 days, and that was the basis for designing the Study 3.

Now, of course, up to the Study 3, the method seems to suggest that there is a potential benefit up to 21 days, because that was the primary endpoint, so Professor Gibbons went back to do the analysis based on the same methods, and that seemed to suggest, also again suggests, strongly suggests that there is a potential benefit up to 21 days.

DR. HIATT: Right. Now, let's just clarify and stay on Study No. 2. Any additional questions for Dr. Hung?

DR. FLACK: Yes. Is it fair to say that the inclusion of the quadratic term was methodologically correct based on the outcome of the study, which was not known in advance? And if that were the case, and it is methodologically

correct, then, why wouldn't you include it if it fits the analysis of the data when it was not known a priori?

DR. HUNG: The model, just for Study 2, for 56 days, the model clearly is not linear. So, therefore, inclusion of the quadratic term makes sense, but because of the presence of the quadratic terms, the rate of change is not on good slope like in the linear model. In the linear model, it really change the slope.

DR. STOCKBRIDGE: Its fundamental problem isn't that. I mean it's true that you can't interpret the linear component of a quadratic equation as representing the slope, but the fundamental problem is that the quadratic term was not prespecified, and it wasn't the hypothesis-generating analysis that was done in Study 1, which was linear.

DR. FLACK: I guess my point here is if it wasn't possible to know the structure of the data beforehand, then, how can you really be critical for appropriately analyzing it--

DR. HUNG: You could have had--Study 1 suggests a linear.

DR. TEMPLE: You could have had a what if. I mean what they could have done--just to back off a little bit, one runs into this all the time. What you are hearing is there is nothing unreasonable about having a quadratic term, what it wasn't was built into the planned analysis.

Had it been built in, in such a way that said if there is this finding, then, I will have a quadratic term, we wouldn't be arguing about it, but that wasn't done.

At the same time, it isn't crazy. I mean this goes right to how perfectly you have to anticipate everything you do and when is a post-hoc reasonable analysis okay. The trouble we have is all post-hoc analyses are reasonable. Who would do a stupid post-hoc analysis? They always look reasonable.

DR. FLACK: You do in hypertension.

DR. TEMPLE: Well, okay, maybe somebody does one that is really stupid, but they all look

plausible, they all have bases, and that's true of every subset analysis that has ever been done, and a crucial question for us, and it comes up all the time, is what are the limits on that. That is really what Jim is saying.

DR. FLACK: I will just make one final comment. You know, I think for things that are not anticipated, you cannot anticipate before you see the data, that you have to have some flexibility in analyzing it appropriately and to do what is methodologically correct, because I think it was Yogi Berra or Will Rogers, somebody said it is very difficult to make predictions especially about the future.

I think in the first study, they choose the wrong endpoint, and in the second study, it sounds like they tried to do it correctly, and unless there is something methodologically incorrect, then, I have a hard time being critical of that second study.

DR. HUNG: I just want to make clear on the record, it is not really, you know, whether

quadratic is better than linear or not for the Study 2. As I said, if you see the data of the Study 2, obviously, you should fit quadratic polynomial model.

The question is after the model is established, what is the rate of change. The rate of change is not the linear terms. So, that is why I did that analysis presented in the Table R22. That seemed to suggest that there is a benefit up to day 14.

DR. TEMPLE: But, Jim, I don't understand that - if you did an analysis and you thought it was appropriate, and you couldn't translate it into a linear description, that would be another way of saying the drug works, but I am having a little trouble describing exactly how it works.

For better or worse, we deal with that all the time. You know, we calculate hazard ratios and then give the median, I mean we always do that, because no one knows, no one understands hazard ratios.

I think the fundamental question is

whether this not anticipated analysis leaves this with a study that actually is achieved significance or not based on our usual standard or a reasonable standard of what the prior planned analysis was.

What I hear everybody saying is this is sort of close, it is not crazy to have done that, not even silly to have done that, but it wasn't exactly anticipated, and I think what you are saying is a very good question.

We worry about slippery slopes, because they always look reasonable. Anyway, that is what is at the nub of Study 2, I think.

DR. GIBBONS: I think there is one issue that is important from a regulatory perspective, and that is, you talked about two confirmatory pivotal studies. You have to understand that at the time, the second study was going to be the second pivotal study.

Now, this may not be the way you are thinking about it, but this was the way it was described to us, because we were told that the second pivotal study would have confirmed the

secondary endpoint from the first study, so it was designed as a confirmatory study.

So, because there was this issue, that is what led to the third study, but the second study was being viewed as the second pivotal study, and, in fact, the MHRA, on the basis of those two studies, one which was a secondary endpoint, and the other where it was a primary endpoint with the quadratic term, approved the drug for marketing in that country.

DR. TEMPLE: This actually goes to another--you are hearing them all--another agony we have all the time when if you do a study that fails, find an appropriate subset or appropriate analysis, appropriate endpoint, whatever, that works, and then do another study, do you have two studies or one and a half or one and a quarter, or something like that, and what is the level of evidence.

If it is a mortality study, we are concerned and then go with it, but on symptomatic studies, it is more common for us to say, well, no,

you really need two studies. Of course, they have done a third study, so we get to answer that, but it is not quite the same as having two independent studies where, you know, if you calculated an overall p-value, it would be 0.00125 or something like that, as having a hypothesis-generating study, and then confirming it.

That confirms it as a level of 0.05, which might be good enough for some circumstances, but it is not quite the same as having two independent studies.

DR. WARNER-STEVENSON: I just want to approach this from another side of this retrospective change. It looked like in the second study, patients just got better faster than we thought they would in both groups.

I want to ask both Dr. Hung and Dr. Gibbons, if you had used the prespecified analysis, but truncated it at 21 days, would it have been a positive trial? I recognize again it's post hoc, but I am just trying to understand the form of this data and the time course.

DR. HUNG: I did not do the analysis, but probably yes.

DR. GIBBONS: I did the analysis, and the answer is yes.

DR. HUNG: Probably yes.

DR. GIBBONS: Clearly. And that speaks to the issue of slope, too. Dr. Hung is completely right, the linear term and certainly the treatment by linear time interaction is not a slope over the entire course of time.

The presence of that, when there is a quadratic term, when there is a quadratic term, it means that it is a piecewise slope. There are many slopes. The slope depends on time. It is not going to be the same over the entire course, so that is why it is important to look at the data.

That is certainly what we did. We looked at the data and saw that, in fact, up through day maybe 45 or 50, the difference between treatments was, in fact, linear. It was only after day 50, just at the very tail, where the difference started to change, but things started to get a little

curvilinear in both groups after day 21, which is why we all agreed upon the third study being a 21-day study.

DR. HIATT: So, just to clarify, you went back to linear, not quadratic for the third study.

DR. GIBBONS: That is correct.

DR. HIATT: Consistent with how you did your secondary analysis of the first failed study.

DR. GIBBONS: That's correct. Then, we went back to both Studies 1 and 2, and redid the analysis using--

DR. TEMPLE: It was also a 21-day study, so the quadratic period didn't come up.

DR. HIATT: Right, got rid of that problem.

DR. GIBBONS: But let's add to that. We did do 56 days of follow-up in the third study, and we prespecified the analysis of 56 days including the quadratic term, which was approved.

DR. LINCOFF: I don't disagree that the 21-day and the slope, et cetera, but I think we need to remember that the fundamental point here is

that you can't let your analysis be driven by the data. I mean that's like changing your endpoint, saying your primary endpoint is a composite and now you saw a difference, because I mean that is invalid, that's introducing another set of randomness, that is, the randomness of a population picking the endpoint or, in this case, picking the analysis.

I think you prespecified an analysis plan, and you didn't use it. That doesn't mean that the data is invalid, but I don't think we can say that it's irrelevant and that it's okay to take the data and let it dictate which analysis should be done. I think that is a very fundamental point that we can't overlook.

Totality data may still support that this drug has effectiveness, but I don't think it's on a basis of you hit a prespecified endpoint by changing the analysis in response to the data.

DR. GIBBONS: And, indeed, that was the reason for doing the third study.

DR. DeMETS: I have a question for Jim or

Bob. First of all, I think to some extent, we teach students that the data should suggest the analyses. You want to fit a model that makes sense. You wouldn't expect me to bring to you an analysis which the model didn't fit the data at all, you would send me back home.

So, you want the analysis to reflect the data as best you can, however, you don't want to start changing the question or the endpoint, or a whole bunch of other things, but to say that I specified the linear analysis and it didn't turn out to be a linear curve and I am going to stick with it come hell or high water would be stupid. I don't think Jim Hung would disagree with that.

You know, you want the data to reflect it, but the question is what is the question, and I think implied, but not specified very well is when you say it's a linear model, I want to compare the two curves or the two equations, two lines.

Well, when a linear model is sloped, it says it, but there is other ways to test that question. You can use likelihood ratio tests, and

so forth, and so on. You can do the same thing for a non-linear curve. That wasn't the way it was set up. So, you could do a likelihood ratio test and say are these two curves different or not. That's not the way the question was posed. It was posed in a simplistic way, and therein lies the trouble that Dr. Temple raised to us.

But I don't think we want to get trapped to be a slave to the analysis plan to the point where it looks stupid. We really need to reflect the data, but you can't change the question, you can't change the outcome measures, and all that kind of stuff, but I think we have no problem with them, but the analysis somehow has to have the flexibility to fit the data we get, but we weren't collectively smart enough in this case to figure out how to pose the question to get around the question that Jim Hung is raising.

DR. STOCKBRIDGE: But you did have 56 days in Study 1.

DR. DeMETS: True.

DR. STOCKBRIDGE: If you thought it needed

a quadratic term, you could have specified that upfront. It wasn't as if the first chance to detect an effect over 56 days occurred in the second study.

DR. GIBBONS: But the data in Study 1 were beautifully linear through 56 days, exquisitely linear through 56 days. There wasn't a hint of curvilinearity either in the curves and certainly not in the difference between the two groups through 56 days in Study 1.

If I had seen such a difference, I would have definitely put in a higher order term.

DR. STOCKBRIDGE: So, now, Dave, what do you do? You have seen two sets of data, one set looks linear, the new set looks curvilinear, how do I interpret a p-value now?

DR. DeMETS: We don't know, and that's the problem. I mean I think expecting a response curve to be linear over 56 days, from my limited experience, would be pushing it, to begin with, and so you might get lucky in the first trial, but 21 days or a shorter period of time, a linear

approximation, we do that all the time. Even though we know the response isn't linear, we take shorter periods of time and say for that period of time, the linearity is pretty good.

So, I think as it unfolded, we got to the third study, it was reasonable to postulate on that third study in linearity and make it for 21 days. I think that was very reasonable.

The problem is the effect isn't as big in that third study, so it didn't turn out as we thought it would be, but I think that the path you followed made sense, and can be sold on non-linearity over 56 days, so let's look at the first 21 days, pose a linearity question, makes perfectly good sense.

DR. TEMPLE: Dave, you said that one shouldn't be foolishly slavish to your initial plan if the data don't fit it. It is hard to argue with that, but you also said that if your original plan was sort of comparing slopes, and you have to do something that makes it not a slope analysis anymore to intelligently deal with your data, then,

you can't quite do that either, because you have changed the question. I am not sure how those two things fit.

DR. DeMETS: If I had said I wanted to do a likelihood ratio test to see are these two curves, even though one may be a straight line, different, and that that's the same whether the equation is a linear curve or a non-linear curve, that's the same statistical test.

Now, it turns out the linearity is essentially the same thing as asking the slope question, so that they come out the same.

DR. TEMPLE: Right, but if they didn't think to put a likelihood ratio question--

DR. DeMETS: I understand, that's right.

DR. TEMPLE: And the expected linear data based on the first study aren't linear, does that mean it is now irretrievable despite your wish that we not stupidly stick to the plan, or what?

DR. DeMETS: I have no answer to that question.

DR. TEMPLE: That is what I was afraid of.

DR. GIBBONS: I think an important factor in that is that we didn't change the endpoint, we didn't change the model. We prespecified a treatment by linear time interaction as the probability value to live by or die by, period, and we didn't change that.

What we did is we added an additional term in the model to capture the curvature that we observed in the second study, but we didn't use inference based on that curvature. We just said is there any difference between treated and control subjects in terms of the linear component of this curve, which was prespecified. We didn't change that. We could have changed that. I could have used the likelihood ratio chi square statistic to get a composite test of both the linear and quadratic effect.

I don't think that is appropriate. I think that is changing the endpoint, but by stipulating it is just the treatment by linear time interaction, that was prespecified. The fact that I add a little curvature is not changing the

question.

DR. HIATT: Dr. Hung, thank you for standing there. Let's just clarify your position on Study 3. So, the sponsor, using what the FDA had asked for, an LOCF analysis, still comes up with a positive, and depending upon how you handle a couple of missing data points, it can become a negative.

Could you clarify your position on that one?

DR. HUNG: I did a lot of sensitivity analysis that's presented in that particular table. The p-values can range from 0.03 to 0.15 depending on how you handle the missing observations.

Now, of course, how to handle the missing values is not strictly statistical because, for instance, the post-discontinuation data, whether those should be counted or not, should be included in the analysis or not is not--I mean after patient dropout, that is not purely statistical.

So, if I take all this range into consideration, I can only say that this study is at

most or at best borderline significance. What I am saying is that 0.05, 0.07, that kind of range.

DR. HIATT: I just wanted the committee to make sure that we are clear on that. I mean ultimately, we have to deliberate, I still think on a standard that has been set before, which is two positive trials for approval or something like that.

Does anybody have any questions about whether Study 3 is positive or negative?

DR. TEMPLE: Just one point. Anytime, not to state the obvious, a study is nominally significant at 0.05, anything you do to it, take one patient away, and it won't be anymore probably, or it has a good chance of not being.

So, I think what Jim says is that it is at the margin. If you have two studies nicely at the margin, and you multiply their p-values, you get considerable reassurance that the drug has the effect.

So, the big question here is what do you do with Studies 1 and 2, can you count them, can

you count part of them, and so on. That is why we have you here, because we already know we are in trouble, right?

DR. FLACK: Just one comment. I know in my world of hypertension, when you do trials, you basically get a range of effects. I mean sometimes you see a big effect, sometimes you see a smaller effect, and if you don't have publication bias and report everything, sometimes you may even see a reversal of the effect just by chance alone.

So, I don't know if there is going to be a logical way to explain the differences in the effect sizes seen across these trials, and it may just simply be that is just what happened when you maybe don't have a huge effect and all you see are variable effect sizes.

You know, looking across these trials, I think that one thing I would say is that these guys didn't predict the future very well, because in the first study they had the wrong endpoint and all, but does that totally invalidate seeing consistent differences in pain just because you weren't really

intuitive enough or insightful enough to actually specify that.

I also note, too, that government agencies, as well as investigators, take great flexibility with data once it comes out, but here, I don't necessarily think that I have heard any gross violations of anything methodological, so I just don't know what to make of the effect size differences except to say that it is probably not inconsistent with just doing trials and all, and reporting data out even when there is an effect.

DR. KASKEL: Can I say something? I am a PI on a clinical trial from the NIH now, and we have gone through a lot of growing pains in the last three years. Our endpoints haven't changed, but we had obstacles that are similar to what you are saying, and expectations that didn't come to fruition, and had the modified inclusion criteria, modify certain characteristics that we weren't sure of, and the effect of the treatment.

So, it's a learning curve and this is an area out of my expertise here, but I think in due

respect to trying to put together a clinical trial, you always have to have the room to make the changes as the data comes in, so you may not intentionally have seen the future all the way--it was Yogi Berra--but you can change.

DR. TEERLINK: Two things. I did actually talk to the medical reviewer in terms of the Study 3, in terms of the U.S. study endpoints at 21 days. In fact, the Study 3 was not positive in Germany, not positive in the U.S., not positive in Israel, and borderline, a 2 millimeter difference in Russia, and the results are largely driven in this small study in a subgroup analysis with all the appropriate caveats by Serbia.

In fact, placebo was better than nitroglycerin in this Study 3, so there did seem to be some change in terms of treatment effects, and whether that is just random walk or whatever, it was clearly not a significant treatment here in the U.S. in Study 3.

You presented the results of Study 1 and Study 2 in terms of percent improvement, and that

is sort of helpful. Do you have actually the actual numbers in terms of the--you know, we have mean average daily pain score at time for the groups graphed out, so we could see those?

DR. GIBBONS: They certainly were presented in the final analysis. I don't know if we have them--

DR. TEERLINK: Actually, I see mostly percent improvement throughout this. Maybe I am missing it, and it's possible. But I see the Figure 7 is percent improvement, Figure 8 is percent improvement, Figure 9 is percent improvement.

DR. GIBBONS: Well, here is an example. This is the subjects with moderate to severe pain from the various studies.

DR. TEERLINK: Right, and that is a selected subgroup. I was just looking at kind of the overall, because we are trying to look at what the overall study shows.

DR. GIBBONS: Sure. This is an example of one. There may be, I am not sure, I am not sure

that we have--here is another one. This is, well, this is again in the analgesic users, so I don't have them for what you are asking for right now, but certainly they could be prepared.

DR. TEERLINK: That is your actual endpoint, right?

DR. GIBBONS: That's correct.

DR. TEERLINK: So, it would probably be useful to see the primary data upon which the actual endpoint is based.

DR. GIBBONS: Sure.

DR. HIATT: I wonder if we are at the point where we should do the last presentation. We are going to be up against lunch. There is a risk/benefit presentation to come, and maybe we should do that and finish up any final questions to the sponsor.

Then, Bob, you had a short presentation, as well. Do you want to do that in the afternoon, or do you want to do that--

DR. TEMPLE: This afternoon.

DR. HIATT: Would it be all right then?

We will probably have more questions later. You have been great.

DR. GIBBONS: I was really looking forward to spending more time up here.

DR. HIATT: Okay.

Risk/Benefit

DR. LUND: This is actually the last but one presentation. There is one after this, but I will try not to detain you for too long.

[Slide.]

My name is Jon Lund. I am a colorectal surgeon from the UK, and I have been asked to come to talk today because I have had the opportunity to use this product clinically.

[Slide.]

I am also going to talk to you today about risk/benefit and hopefully, by the end of this short presentation, persuade you that the benefits of Cellegesic far outweigh any potential risks.

[Slide.]

We have talked a lot about this and hopefully, have persuaded you that one Phase 1 and

three of the Phase 3 studies provide the evidence that nitroglycerin ointment applied intra-anally accelerates the pain relief associated with a chronic anal fissure.

Healing in these studies was about two-thirds of the same as many other studies in the literature, although I will say again that healing is not a prerequisite for pain relief.

[Slide.]

We have talked a lot about headache also.

Rectogesic, which is the trade name of Cellegesic outside the U.S.A., have been approved for some time now in Australia, New Zealand, Singapore, and South Korea, and almost a quarter of a million tubes of Rectogesic have been sold in Australia. After those quarter-million, only 10 complaints of headache have been reported to Cellegy.

Since May of last year, Rectogesic has been approved not only in the United Kingdom, but also in 19 countries of the European Union.

Rectogesic has a black triangle. I just

want to make it very clear that this is in no way equivalent to the black box, the black spot I was going to say, the black box in the U.S. This is a device which is attached to all the medicines and medical products to encourage reporting of adverse events in those people prescribing it.

Despite this encouragement, and more than 34,000 tubes being sold, only one report each of nausea and dizziness have reached the authorities, and no reports of any headache.

[Slide.]

For many years, the standard treatment for anal fissure was surgery. There is no doubt that surgery is extremely good at relieving the pain associated with anal fissure and curing anal fissure, however, surgery does have its drawbacks.

We have heard earlier that lateral internal sphincterotomy results in the impairment of continence in up to 35 percent of patients having this operation.

This data here on the graph is taken from the UK Government Department of Health figures. It

shows a time period between '98 and the present day. The '98 guidelines were issued and NTG ointment became more available in the UK.

You can see over that time period that the total number of anal surgeries, less procedures for anal fissure remained just about constant, and also remaining just about constant is the number of diagnoses of anal fissure. What has changed is the number of operations performed for anal fissure, which is halved over that same time period.

[Slide.]

Anal fissure affects young to middle-aged adults. The pain described by these people and defecation is consistently phrased as it's like passing broken glass. It is a very nasty thing to have on defecation, and it significantly affects the quality of life of young to middle-aged people. It stops them from going about their normal business and it stops them from going to work.

[Slide.]

Study 2 includes a Gastrointestinal Questionnaire, which has a few questions associated

to fissure symptoms, but results favor the subjects having improved quality of life, and it didn't suggest that frequency or severity of headache had detrimental effects on quality of life.

[Slide.]

This study by Nira Griffin from our group in Nottingham also looked at quality of life of people with anal fissure. The most significant determinant of poor quality of life was pain, the worse the pain, the worse the quality of life. Nira used this using the SF36. This is a well-known quality of life score.

If you look at that across each of the domains of quality of life, which were examined in the SF36, you can see the higher levels of pain associated with significantly higher body pain and significantly poorer general and mental health, significantly less vitality, significantly decreased physical and social functioning, and greater limitations due to physical and emotional problems.

So, across all those things examined by

the SF36, making up quality of life, pain from anal fissure had a significantly detrimental effect. So, it seems crucial that anything we can do to accelerate the relief of pain in these patients is essential in getting them back to normal activities.

[Slide.]

I was quite shocked to see the data presented earlier about the extemporaneously compounded nitroglycerin ointment in the States, and I was very surprised to see that almost half of the retail pharmacies didn't meet USP criteria.

We, before the introduction of Rectogesic in the UK, had a very similar system, GTN, as it is known over there, was made up in manufacturing pharmacies, and I like to think that when I prescribe a patient something, that that is the thing that the patient gets, but looking at this study, there was over 100 percent variability in the potency of the NTG ointments, and even from the same pharmacy, there was no guarantee that you get the same product every time.

I think, as Dr. Abel said earlier, that one of the reassuring things about having a branded preparation on the market is that you know that what you prescribe for the patient is exactly what the patient gets.

[Slide.]

I have used GTN or NTG since 1994. In Nottingham, we did some of the early studies on this, and we reproduced the first prospective randomized trial which was responsible for wrecking Cellegy's Study No. 1.

Over that time, I found topical nitrates do work extremely well in patients with anal fissure both in terms of pain and healing. Since the introduction of Rectogesic, which is exactly the same as Cellegesic, my colleagues in primary care have been very enthusiastic about taking it up. As I have reported, it works very well in primary care also.

This is very important. The patient doesn't have to wait to be referred to secondary care before they can start effective treatment, so

when the patient arrives, they get a script for the Rectogesic, take it and apply it, and the pain relief happens very quickly rather than waiting an extra week, or a month, or six weeks as in the UK to see somebody in secondary and for treatment to begin.

Headaches do occur, of course, but in my experience, and I have treated a lot of people now with topical analgesics, and it is very, very unusual for patients to stop treatment because of headaches, and this is because the pain is of a completely different order of magnitude.

People do get a headache, but it is easily treated with simple analgesia, but the pain, as I said before, of anal fissure on defecation is like passing broken glass, so patients are unwilling to put up with that pain, and willing to put up with a headache to relieve that pain.

Also, in our practice, since the use of topical nitrates has increased, and particularly after the introduction of Rectogesic, we have noticed few referrals from primary case, so these

patients with fissure are being managed in the community rather than come into the hospital, which saves money for the NHS, and operations for anal fissure have been fewer in line with the Department of Health statistics I showed you earlier.

[Slide.]

So, in conclusion, nitroglycerin ointments provided as a GMP product will assure accurate dosing, something which I and I think my patients will find reassuring.

The benefit of accelerating the rate of pain relief and potentially decreasing the need for surgery, by use of Cellegesic, clearly outweighs any risk of adverse outcomes.

Thanks.

DR. HIATT: Just to clarify one of your last statements, that headache is rarely a cause of discontinuation in clinical practice--

DR. LUND: Indeed.

DR. HIATT: But there are clearly several tables in here that show the dose response between dose of drug and headache, and had withdrawal due

to headache, would you agree with that?

DR. LUND: In the studies it may be, but in clinical practice, it is very unusual for people to stop treatment because of headache. It does happen from time to time, but it is infrequent.

DR. HIATT: Just to make the point that they are in here, as well. Then, you were talking about risk and benefit. Just to state the obvious, we haven't seen a safety database here. Just to clarify that it's a young, relatively healthy cohort of people, there aren't a lot of people that are older and have cardiovascular disease, so we haven't seen a lot of safety data other than tolerability.

I guess at some point, the group should talk a little bit about are there subgroups that we might be concerned about where risk could be an issue. I think it would be very hard to ever quantify that, because the event rates around that risk would be extremely small, but we shouldn't forget that concern.

DR. HARRINGTON: I have two questions, one

on data and one on your clinical experience. Let me do the clinical experience first. How long in your practice would you typically treat someone with this therapy before they get relief, and how soon after they get relief do you stop the therapy?

DR. LUND: The relief will be obtained in the first few days to two weeks. We will continue the ointment for six to eight weeks and then review them at that time.

DR. HARRINGTON: My second question has to do with the Griffin study. How many patients were in that study, and secondly, you make the statement that the pain assessed by the VAS correlates well with the SF36, and this has been something, as you have heard all day that we are grappling with, is what magnitude of change in the VAS is clinically important. Does that come out in the Griffin study?

DR. LUND: There is 54 patients. I can't remember the magnitude.

DR. HARRINGTON: One of the things that we have been grappling with all morning is how big a

difference in this population in the VAS. We have heard that in the dyspnea world, a change of 20 is clinically meaningful, and I trying to get a sense of what is clinically meaningful in this world.

DR. LUND: I am not sure that I know the answer to that. I think all these VAS scores are quite subjective, and may be subjective from person to person, as well. People fill in scores differently. Your endpoint may be for the pain you experience rate different to the next person, all depending how vivid your imagination is about what the pain might be.

DR. HARRINGTON: So, maybe I could follow up with a different sort of question. This morning we heard from an American surgeon that prior to using nitroglycerin, he was operating on about 60 to 70 percent of the patients that were referred to him.

Are those numbers similar in your practice or in your experience, and then, secondly, what is your rate of operation today?

DR. LUND: The situation in the UK is a

little bit different to that in the U.S., and because there is a gap between being seen in primary care and being seen in secondary care. The patients that we see are always well- established chronic anal fissures. These tend not to get better by themselves.

Our operation rate would approach 100 percent before this. Since we started using topical nitrates, two-thirds of these are healed, so we don't operate on those, and then a third proportion of people will elect to continue topical treatment rather than have surgery.

DR. HIATT: If it's all right, maybe we could go the last one. Did you want to ask a question?

DR. PORTMAN: Yes, I am concerned particularly, as Bill was saying, about the elderly and those who might have cardiovascular risk. The data presented earlier, looking at blood pressure, where there is 10 percent of patients had a drop of 20 millimeters of mercury diastolic, it was concerning and yet the placebo had the same, which

makes me wonder about how well the blood pressure was really being measured.

Do you have any data on blood pressure from your clinical experience?

DR. LUND: No, we didn't measure blood pressure in any of our trials and certainly not other than our clinical practice. It is not a common condition in the elderly. It is by far, the majority of patients are young to middle-aged adults, but I have never known, nor heard of any hypotension-related complication of administering topical nitrates.

DR. HIATT: Maybe if we could go to the concluding comments in the interest of finishing the morning, and the afternoon devoted to some more discussion.

Summary and Conclusions

[Slide.]

DR. GARVEY: I am Tom Garvey. I am a gastroenterologist and I am a consultant to Cellegy over several years, I used to be the supervisory medical officer in the Cardio-Renal Division a long

time ago.

Most of what I was going to say is irrelevant. Gratifyingly, this discussion has moved very, very rapidly beyond the point what I was going to say is useful.

I think the issues here, the main issue is missing data, pain measurement, and other major issues are well engaged at this point.

Rather than persist in comment, I do want to make one observation, that much of what has been discussed here this morning has been traversed previously in interactions between Cellegy and FDA during the special protocol assessment of which you heard and which resulted in the protocol for the third trial, and the agreement was that at least implicitly, that if this trial was a success, the drug was likely to be approvable.

I will leave, at this point I will desist and leave you to grapple with that problem.

DR. HIATT: Well, thank you.

DR. TEMPLE: Can I ask one question? It could be of anybody. I hadn't fully appreciated

this, but in Dr. Marciniak's review, it is clear in Study 3 that the entire effect of the drug comes from the Serbian subset, all the others go the other way.

I just wondered if anybody has any comments on that. I mean Germany goes the wrong way, the U.S. goes the wrong way, Russia goes the wrong way or sort of neutral. It's all driven by the Serbian subset.

DR. FLACK: Analogous in the hypertension, diabetes, renal world, is some of the diabetic nephropathy trials, have seen some quasi-similar, not exactly, where you look, for example, with RENNAL study, basically, the only group you saw benefit in really was the group in China, in Asians, and it is probably because they had the highest level of proteinuria.

Now, it didn't go the opposite direction, it was just neutral in U.S. in Hispanics. It is probably a little hard to explain the opposite direction unless you are really not necessarily looking at a very large effect where you might say

that you might expect that some of the places would go the other way if you didn't have a real huge effect, and I don't suspect that this is a real huge effect.

DR. TEMPLE: We were actually quite troubled by that observation in both RENNAL and IDNT, however, that applied to the measurement of creatinine doubling. When you got to more tangible endpoints, end-stage renal disease, actually, the effect looked pretty consistent across all regions, which reassured us in that case. Of course, we don't have anything quite analogous here.

DR. HIATT: To follow up on that question, I don't remember, was a treatment by country interaction looked at in Study 3?

DR. GIBBONS: No, I don't believe there was a treatment by center interaction. Center was included, of course.

DR. HIATT: The center usually is, but this is different, this is country.

DR. GIBBONS: I think what is important to note about that is that that was an endpoint

analysis. I mean obviously, we are not powered to detect these kinds of differences, and the analysis was not based on--it was trying to look at effect sizes, but they were looking at effect sizes using last observation carried forward endpoint analyses.

Different results could well be obtained when looking at all of the available data over 21 days rather than just the final point, particularly when many of those differences were quite small. They might have been quite small at 21 days.

We also know that 15 days was the point of maximal effect of the drug. You could have come up with a very different interpretation at 15 days. I haven't done those analyses, but just to put it into the context.

DR. HIATT: I guess, I don't know, there have been other examples in cardiovascular trials where there have been positive treatment by country interactions where drugs seemed to work pretty well in Europe, but didn't work too well in the U.S., at least in my peripheral vascular world, where the overall p-value was strongly positive for the

overall result, which is what you would want to most likely to believe, but it just detracted a little bit from the overall cleanness of whether the study was cleanly positive or not.

DR. TEMPLE: We certainly don't have unequivocal policies on how to deal with those things. To some extent, you are reassured if it's a survival or hard endpoint study, some of the things you might worry about seem unlikely in that setting. These are all symptomatic conclusions, and I think to be candid, you worry more.

This is all on page 57 of Dr. Marciniak's review. In Germany, this is absolutely true, this is just the 21-day score, so we could take a look at some other kinds of analyses, but the German and U.S. data just go plainly the wrong way. The Serbian data are where the best results are, and Russia is sort of neutral, but that is just on 21 days, that is absolutely right, and it is very hard to know what to do with those things.

We encounter them a lot. Multiple studies help, but it just seems worth pointing it out.

DR. GIBBONS: One point about the U.S. data, it is becoming very difficult to recruit subjects in the U.S. for doing that, because the drug has become so widely available, so recruitment in the U.S. was very, very down relative to the previous two studies.

DR. HARRINGTON: Bob, I an curious, because certainly, in the large cardiovascular trials, that I am most familiar with, we frequently see, as you know, geographic variations, and we look at the demographics, we look at the treatment parameters, it is bothersome when it is not a hard clinical outcome that the region stands out.

Did you begin to do analyses of trying to understand the population in Serbia relative to the other regions, what the background treatment was relative to the other regions? Was there something else that stood out?

DR. TEMPLE: No. As everybody knows, we are seeing more and more non-U.S. data in more and more settings, and we are just paying attention to see if there is something there we should be

looking at.

I wouldn't pretend that we have the answer, but some of the large cardiovascular trials have had differences. The metoprolol trial is certainly one of them, but while mortality was different from one part of the world to the other, the overall benefit was present in all regions, and that reassured us, so we looked for all of those things.

The International Conference on Harmonization guidance does say that a region has the capacity, it doesn't have to exercise it, to insist that there be replicated findings in its own region in case they are nervous, but it is hard to know whether to be nervous.

I mean I know all the papers on what is wrong with subset analyses, I quote them all the time, but it doesn't mean you don't notice.

DR. HIATT: I am wondering, given the hour, if we shouldn't maybe break for lunch. Do you want to come back at 1:30, is everyone agreeable to that? Thank you.

(Whereupon, at 12:40 p.m., the proceedings were recessed, to be resumed at 1:30 p.m.)

A F T E R N O O N P R O C E E D I N G S

[1:40 p.m.]

DR. HIATT: We will get started again. I think everyone is back, a little bit late, but we still have time.

You wanted to make a comment?

DR. AZARNOFF: I understand there was a question regarding the quality of data in Serbia. I should point out that the Serbian clinical site was audited by the FDA and they did not get a 482.

DR. HIATT: Okay. I think the next agenda item, Dr. Temple, is a brief presentation.

Presentation

DR. TEMPLE: The title is "Can the Effect be Too Small"?

[Slide.]

This is going to be a brief discussion of one of the questions that is raised by the questions that you will be asked to answer later. There is really two; one, have they shown that there is something, that is one question. The second is, is it possible that there is an effect

size that has been demonstrated, that it is so small a drug shouldn't be approved for having it.

I just wanted to run through a little bit of history. I will tell you at the outset there is nothing definitive anywhere in the law, regulations or our own guidance that gives an unequivocal answer to that question.

So, I will go through what the law says, which is not very much, a little bit about the legislative history that is relevant - one critical court case, a statement we put into the Federal Register and the public from 1996 under the auspices of President Clinton and Vice President Gore that sort of bears on it a little bit, an item from our recent Patient Reported Outcomes document, and then just one thought I have that I already expressed before, about the way we tend to present data as a means, and not pay too much attention to the distribution of results.

[Slide.]

The legal standard for approval is that an application has to include, "substantial evidence."

It is not compelling evidence, a word you will see in your questions, it is substantial evidence is the operative term, and that means, according to the law, evidence from adequate and well-controlled studies.

The word "plural" was, according to the writers, intended, they meant more than one, but a modification of the law in 1997, the FDA Modernization Act, said that in some cases, one study can represent substantial evidence, one study plus confirmatory evidence can represent substantial evidence, and it never said what confirmatory evidence meant.

Now, if you just read the language of the law, it sort of implies that any truthful description of any effect at all would be a basis for approval as long as it's truthful, but there are at least two bases for thinking that's not entirely true. One is that there is a safety requirement, and another is a Court of Appeals case called Warner-Lambert v. Heckler.

[Slide.]

The safety rules say that an application can be rejected if the tests of a drug show it's unsafe or failed to show that it is safe.

We have tended to say through the years--you will see this in slides, if not in any document--that since all drugs have adverse effects, safe must mean that benefits outweigh risks and that any other meaning is illogical.

If you believe that, it suggests that effect size could matter. Certainly, if the drug were very toxic, it might matter, and I think nobody doubts that that is true, but it also might lead you to ask that since all drugs have adverse effects that you don't know about yet, unknown things, rare things that you haven't picked up, could it also mean that an effect size that is just too puny could be outweighed by that unknown risk, and I have nothing more to say but to pose the question, because I don't know of anything that ever has addressed it.

Now, having said what I just said, it is not quite clear that the people who wrote the 1938

law on safety really did mean that benefits have to outweigh risks, because they didn't demand any evidence of benefit, and since they didn't demand it, nobody provided it in any way that we consider meaningful.

So, they may not really have meant that benefit outweighs risk, whatever we now think, but only may have meant--they may have just meant that nothing really too awful was seen, so it is not quite clear. The law doesn't help much, in other words.

[Slide.]

Warner-Lambert v. Heckler, not that old, from 1986, basically said that just because you have shown something doesn't mean you have satisfied the requirements of the Act, that the effect has to be clinically meaningful, and not therapeutically trivial.

They specifically rejected the argument that any effect claimed, if it was supported statistically was sufficient, and that the size of the effect is irrelevant, they rejected that.

On the other hand, the things they were talking about really went to a different question, which was whether the effect that had been shown mattered, so that one of the effects they were looking at was the reduction of fungal load in the gut without any evidence that that led to improved anything, and they didn't think that meant anything.

Going back in history, there used to be drugs that increased bile flow. Well, is that a good thing or a bad thing? So, how much this should be taken as support for the idea that a documented effect, that if large enough would be meaningful, should be rejected is not clear.

[Slide.]

Now, the legislative history at the time of the 1962 amendments went out of its way to try to reassure people who were at the time very, very worried that the standards were going to get so high that no one would bother to develop drugs, and they particularly wanted everybody to know that there is a "no" relative effectiveness requirement.

Now, relative effectiveness usually means you don't have to compare a new drug to the available therapy, so they very clearly said that a new drug doesn't have to be better than or even as good as available therapy, and as the quoted text there says, they wanted to make the point that they had struck a balance between the need for some assurance that new drugs are okay, and that they are not placed on the market until they have passed appropriate tests, but also, a simultaneous need to assure that government control doesn't become so rigid that the flow in the drugs to the market and the incentives become stifled.

They never actually said, of course, those words don't say that any effect no matter how small is sufficient, but they clearly had some of these concerns in mind, and they wanted to make it clear that a drug didn't have to be particularly effective or even as effective as other therapy.

[Slide.]

In 1995, apparently reacting to concerns--I am not sure where these came from--that

FDA was imposing new standards for comparative studies, which we were not doing then and are not doing now, although you hear in editorials and other places arguments that we should insist on comparative data--that is certainly a popular view that is in some places--anyway, they wanted to reassure everybody that FDA weighs a product's effectiveness against its risks, and considers such things as seriousness of the disease and alternative therapy, but do not require new drugs to be more effective than existing therapies, nor necessarily require a comparison with other products.

Now, the "necessarily" meant that where a product has some ability to prevent a life-threatening disease, or prevent irreversible morbidity, or treat a contagious disease that would really be bad, then, it is essential for public health protection that a new therapy be about as effective as existing approved therapies.

They actually said that the new therapy has to be as effective. They, of course, didn't

mean that. The only way to show that would be to be better. What they really meant was that it needs some appropriate non-inferiority standard, but they didn't get into that at the time, and it's a concept that has been difficult for a lot of people for a long time, so we can't blame them too much.

That doesn't quite say that we don't care about effect size. It does say that you are not in the business of making comparisons.

[Slide.]

So, I think it is clear that we are prepared, and have been prepared in the past, to say that an effect is clinically meaningless, but it has been much more likely that we would say that, because we didn't know whether the effect translated into anything useful, the obvious older cases or increased bile flow, suppression of gut fungus, but you could make the same case about some kind of surrogate endpoint that you didn't think was well enough established, you know, yes, you have done it, but, no, we don't know what that

means.

As a general matter, though, for non-toxic drug and for a not serious disease, we generally have not demanded an effect of a particular size or required comparisons with other treatments, and that doesn't mean we don't worry about whether studies are getting so big they can now detect trivial things, but we haven't done anything to say you mustn't do that, not in a systematic way.

There have been some local places where people, where we actually have said this trial is too big, you are going to be able to detect something too trivial, but not in any systematic way.

We clearly are prepared to conclude that a small effect is outweighed by toxicity. As an example, we have rejected at least two Alzheimer's drugs, one because it caused severe nausea and vomiting, the other because it caused truncal proximal weakness, and there are many, many other examples of drugs where toxicity barred approval, but in those cases, that was because the benefit

plainly didn't outweigh the risks.

In at least a couple of cases, as indicated in the next one, we have found drugs that were too toxic for general use to be approvable if they showed something special, so that, for example, clozapine, a very effective anti-psychotic, was approved on the basis of showing that it worked in people who had failed on other anti-psychotic therapy, and bepridil, a drug which causes torsades de pointes, and still causes some fatal cases even now, was approved when it was shown to work in diltiazem nonresponders.

So, in those cases, you do need comparative data and implicitly at least, that is another way of saying the effect size has to be fairly substantial.

[Slide.]

When a disease is serious and there is existing therapy, we always get comparative data because it is the only ethical study you can do. The only thing you can do in that case is a non-inferiority study.

In those cases, we generally insist on preserving some fraction of the effect, and it represents a compromise. If you are required to preserve much more than 50 percent of the effect, in most cases you get a study that is too large to do.

Now, the exception to that is antibiotics where the effect size is so large, that you really can show that you have preserved most of it and still have a doable study, so in cases like that, you probably have to rule out a 10 or 15 percent difference, and that is what you do. That is considered clinically not so much of a problem.

In reality, it really means the new drug has to be about as effective on a point estimate basis as the previous drug.

[Slide.]

The document we have recently put out for comment, a guidance document on patient-reported outcomes was particularly concerned that these methods are so sensitive, they are creating a new ability to discover effect sizes that may not

matter.

So, it specifically says it is important to consider whether the detected changes are meaningful, and it calls for in advance of the study, specifying a minimum important difference, MID, as a benchmark for interpreting mean differences.

That is clearly the most explicit statement I have been able to find of the idea that a statistically significant effect on a valid measure might not be accepted as evidence of effectiveness, because it was really too small.

An important question which we are discussing internally is do we really mean that, are we going to start to say not good enough, because the effect size is too small, and why would that be true if it is only for patient-reported outcomes, and we have not discussed that in any length. I am not going to tell you what the answer is because we don't have it yet.

[Slide.]

A couple of things are worth thinking

about. One, I mean there is more and more interest these days in individual responses. We were happy with means for a long time, but now that people believe that there are likely to be subsets of the population that respond differently, there is much, much more interest in seeing what the individual responses are.

While we tend to look at mean effects, it is perfectly obvious that individuals will, in fact, have a range of effects, some larger, some smaller, and, of course, what we are most interested in is whether there is a subset of the population in which the drug works like gangbusters, or alternatively, doesn't work at all.

It may be that we are missing things when we focus on the mean. So, one question that we are raising internally, and I am just putting out here, is whether we more often should show both the mean effect and the distribution of effects where that is common.

We have done that for a few classes of drugs, notably, the Alzheimer's drugs, at least

partly to show that there weren't any really huge effects, but it does show the distribution, and what pretty much invariably happens, although obviously, there could be an exception, is that if you look at a cumulative distribution, and you have won on the mean effect, you will always see the distribution curve shifted toward the advantageous side, usually, more or less consistently across all levels of effect.

So, I don't think it represents different data, it is really just a different way of displaying it, and one of the things that we are going to need to think about is whether if you win on the mean, or whatever your primary analysis is, it then would be okay to just show those distribution results without worrying too much about whether that was a planned effect, and whether breaking it down into quintiles is a better way, that all needs some discussion.

Then, of course, the crucial question is what do we mean by effect size anyway. We tend to focus on the point estimate of the effect. That is

not really quite the right thing to do, because we have designed all our studies to show that the new drug is better than nothing at all, our null hypothesis is that it isn't better than nothing at all, and a treatment is successful if it's likely to be better than nothing at all.

If we were really serious about minimum effect, minimum important difference, we would have to revise our null hypotheses to rule them out. I think the consequence of that in terms of succeeding in studies are really interesting to contemplate, and would make it quite difficult, but I think that is an implication if we really think that an effect has to be of a certain size, we have got to put it into practice, we can't just sort of look at the means and say that is not good enough. That is not really intellectually sound.

So, that is my introduction. I would be glad to answer any questions--oops, sorry, you never know.

It's really related to the same question, do we really want to specify a minimum mean effect,

or some minimum difference on some dichotomous measure like number of people with a 50 percent effect, and, of course, the question on all of these things is how you would support any particular value for either a minimum or the distribution or anything, how would you do that.

Probably the answer is you would ask people what they would consider valuable. That is how patient-reported outcomes are generally developed. But that would represent a considerable new and novel effort.

That is the end, I think.

DR. HIATT: Thank you. Questions?

DR. PICKERING: I have a general question, which is not actually specifically related to what you said, but is something I would like to hear about.

One of the things we heard this morning was that this ointment is being used off label quite extensively, and nitroglycerin is not an over-the-counter drug.

Does this have any legal implications

about what is going to happen, I mean depending on what we decide today or to the off-label usage of this ointment, which we heard is very variable, as well?

DR. TEMPLE: I don't know the actual status of these things that are prepared. There are circumstances in which compounding is considered acceptable, but that is usually for an approved use, and this is not an approved use.

On the other hand, we have no illusions that we capture all the people who are doing these things, so I am not sure what effect having an approved version would do.

It might, in fact, permit compounding to occur legitimately, but I am no compounding maven, and I don't really know the rules. I am not sure it would have an effect on it.

DR. WARNER-STEVENSON: One of the very nice corollaries to your proposal of shifting a little bit more to this is that one could probably generate data in such a format that would be much more understandable by patients, and could, in

fact, be part of labeling, 50 percent of patients improved at this level, and get away from the current heinous practice of perpetuating relative percents into the media and out to patients, which lead to gross overestimate on their part of what effect they are getting.

DR. TEMPLE: One of the things we like about showing the whole distribution of results is it doesn't allow someone to pick up the favorite number that happened to work out. We do see that. Actually, 30 years ago, hypertension studies used to be mostly devoted to showing what fraction of people get to goal. You know, that was a standard endpoint.

We tended to discourage that, because you can increase the number of people who get to goal by lowering the average starting blood pressure, and it is true that comparison with the placebo group doesn't get any better, but the number really looks terrific and sort of exaggerates it, so we moved toward means.

But I don't know, I am personally having

some reservations about that. I think it's informative to show both, and it probably would help explain it to people.

DR. FLACK: An interesting way to look at it. I think this is still going to be as much art as a science ultimately, because you are also going to have to try to factor in what are the available therapies.

In some areas, you are going to be willing probably to accept less when there are multiple options and choices out there, or accept a higher level of risk/benefit or more favorable risk/benefit, and you may be willing to accept a little bit less favorable when there is not much out there.

I support the notion of showing the distributions. We did that in a paper looking at black-white differences in hypertension, showing how the ACE inhibitor data has been grossly overinterpreted for showing racial differences, because all you got is a shift in the central tendency, but the distributions almost entire

overlap.

So, I clearly support that.

DR. HIATT: In the interests to put your comments in context, at least traditionally in this committee, as we were saying earlier, sort of two, Phase 3's, cleanly positive in their primary endpoint, which means they beat placebo in this case, and an adequate safety database is what we typically talk about for approval.

DR. TEMPLE: That is, and for a lot of things it would be hard to know what to specify. I mean I don't know what increase in exercise in an angina trial on a somewhat artificial situation corresponds to something meaningful.

We have just internally been discussing a recent trial in which the average number of angina episodes went down by about one per week, which sounds pretty unimpressive until you start thinking about it, and my thought is you probably have the number of angina episodes that corresponds to the angina-provoking work you do, and most of the time you get angina whether you are on a drug or not,

and only in a few cases will you stop work at just the right time when you didn't quite get an angina, so, I don't know, maybe one is pretty impressive.

Part of the reason we don't stress this is that the situation under study is highly artificial, and it isn't always clear how to translate them directly back, so in a lot of cases we don't. You improve on the measure that we believe is a valid measure, and that's okay, but this case raises the question of whether that is always the right thing to do or what.

DR. HIATT: Just to clarify, you didn't ask this sponsor to achieve a minimally important difference.

DR. TEMPLE: No, we didn't. I don't know of any case where we have done that yet, but the patient-reported outcome document suggests that that might be something to think about. It is presented tentatively and with consciousness of the concerns that are raised by it.

We have also seen in a number of cases that the standard trial in a particular condition

is getting bigger and bigger and bigger, and it raises the question of whether you are now able to detect things that might not matter so much. I mean having raised the question isn't to answer it, but it is a concern that is coming up.

DR. HARRINGTON: I have been asking the question all day about what these changes in the Visual Scale actually mean, but as I see you play it out here, Bob, it does open up another can of worms, doesn't it? I mean in the cardiovascular trials world where minimally important differences now are a well-accepted term, but not a well-defined term.

We used minimally important differences in the non-inferiority and the equivalence trials, but yet there is great consternation over what that minimally important difference actually means, that we are trying to exclude with the upper bound of the confidence interval.

So, I am curious as to, as you have thought about this, and you talk about something even more subjective, patient-reported outcomes,

where is the data going to come from on minimally important differences.

DR. TEMPLE: In non-inferiority studies, the first thing we worry about is not the minimally important difference, it is the effect size the control had in the trial, and that is an important distinction, and if I get off my bottom, I will write to JAMA saying that they completely missed the point in two recent articles where they focused on the clinically important difference and forget about the difference that the drug can be attributed to have, which is the most important thing.

Having established what the effect size is, what you think the effect size of the control agent is in a non-inferiority study, you then ask how much of that effect do we think we need to preserve, and nobody for a minute would say that that is a rational, carefully thought-out document to choice.

It is somebody saying hmm, this drug has a mortality effect, I don't want to lose more than

half of that, and that is about the level of sophistication we are talking about, and the reason you have to do that is you either get no more new drugs of that class, or you have to do studies with 50,000 people in them, and that is not possible, so you end it.

The other thing, of course, which we don't know how to quantify very well is what you rule out at 95 percent is not the only measure in the trial. There is a point estimate that matters, and so when we describe what you have to rule out, we are saying what you have to rule out with the usual level of confidence that you rule out in an effect of zero in a placebo-controlled trial.

But those really don't try to define the minimum important difference, they sort of assume it, and you are always involved in some important endpoint like death or stroke or something like that.

There is very little experience on what minimum important differences might be in symptomatic conditions, and there is no experience

to speak of with non-inferiority studies in symptomatic conditions, because you don't have to do a non-inferiority study, and it is the devil to design one.

DR. WARNER-STEVENSON: I just wanted to mention I think one other related issue here, which is that one could certainly see that if you had a medication, for instance, that decreased in everybody a certain symptom by 3 percent, that that would be trivial. However, if the same medication, the same mean data took 25 percent and made them symptom-free, that could be very important, particularly because in real life, we don't continue something that doesn't work, so we wouldn't be treating the other 75 percent.

DR. TEMPLE: Well, that's the attractiveness of showing the individual results, as well. I have to say, however, that in the experience we have had so far, it is pretty continuous, that a drug that has a little effect tends to have a little advantage on every degree of improvement, usually small, but there could be

exceptions to that.

I would have thought actually ACE inhibitors, in looking at blacks and whites might be different where you get a bunch of very large effects in one group, but you are saying no, so that goes against what we all expected anyway, nothing like the data.

DR. HIATT: We are going to come back to these questions. I think perhaps we should come close to wrapping up our general discussion. I think one thing that the committee has identified, that might be not fully flushed out is the safety side of this compound, and maybe we can kind of go through that rather quickly.

I think the context is a lot known about nitroglycerin, fairly young, healthy cohort of people, certainly a few at the margin that might be older and not tolerate hypotension very well, a short course of therapy, so that the absolute risk that may be unresolved is going to have to be relatively small.

On the other hand, we are not looking at a

fully flushed out safety database that typically, you would look for in a symptomatic endpoint-driven development program where usually there would be several thousand people exposed and events would be counted, and we would have some certainty or some confidence around the margin of those risks.

We don't have that here. We have got tolerability information and a long history with nitroglycerin, and yet in the context of giving people doses that may have different levels of bioavailability, and not being given for people who are actually having symptomatic angina at the time of taking the drug.

We also don't know a lot about drug-drug interactions, particularly PDE5 inhibitors and whether concomitant vasodilator therapy might be an issue, too. I guess what I would like to ask the committee and the sponsors to resolve any final safety concerns, because then we can put that in context with our discussion around approvability around efficacy, so those are just my preliminary thoughts. I will turn that over to you all.

[No response.]

DR. HIATT: Unless you all just agree with that and you want to move on.

Committee Discussion

DR. WARNER-STEVENSON: I thought maybe John would speak up, but I will speak up. The blood pressure remains a concern to us, because the only time it is measured at peak drug effect is that first administration in which, you know, you have a 4-fold higher incidence of diastolic blood pressure falls.

It is small. On the other hand, from a cardiac experience, I am pretty sure most cardiologists have experience of some young person who comes in with atypical chest pain, gets nitroglycerin in the ER, and has a bradycardic asystolic arrest, so it is not trivial even in the young, healthy population although it is very rare.

I don't know what we can add further to that other than a cautionary note.

DR. PORTMAN: I mean that is my concern, too, and that is why I asked the question. We have

all seen perfectly good drugs that work well, that we have not approved because they have, a small proportion have a major side effect. It really worries me that we are using a compound that has known vasoactive effects without really having a good safety database.

What if we approve this and there is 700,000 people in this country, you know, all of a sudden go on this drug and we start to hear reports coming back of these events, that would be very disturbing.

DR. HIATT: So, that is going to remain an uncertainty today. Is there anything, Ron, you would like to pursue in that regard?

DR. PORTMAN: No, I just think it needs to be, you know, perhaps a cautionary note if we decide to approve it to people who are using it.

DR. TEERLINK: I thought you were actually giving the sponsor a chance to present more safety data, so that is why I was so silent, but, yes, I share Dr. Stevenson's concerns, as well.

The other thing that was of interest in

reviewing the AGA document, they expressed concerns suggesting that actually, patients treated with nitrates, topical nitrates, have a higher incidence of rebound ulceration, and I don't see that we have a database that actually can help us evaluate that or not.

You can try to answer it from experience, but I don't think we have randomized data.

DR. FLACK: I have a question for the sponsor. When they actually took blood pressures, did you actually have a blood pressure measurement protocol? One of the problems of taking blood pressures in these trials is if you don't have a blood pressure measurement protocol, the pressures are usually wobbly like crazy, full of error, a lot of terminal digit preference, and it is almost hard to get an accurate picture of anything. It's like trying to take a picture and the camera is moving up and down.

Also, too, do you have the data in its rawest form? Do you have what the blood pressure change was as opposed to just exceeding certain cut

points, and I would be probably more interested in systolic than diastolic.

DR. LUND: I was just going to make the point again that a quarter of a million of these tubes have been sold, and there have been no reports of hypotension-related complications of administration.

So, while it is not randomized data, I think it is beyond my own clinical experience and the clinical experience of several countries around the world, that there has been no adverse events related to hypotension.

DR. TEERLINK: Though I haven't reported hypotension in response to intravenous nitroglycerin ever either.

DR. HIATT: Just to be frank, I am not sure that is very reassuring. I think we only know what drugs do when they are compared to a placebo control in this particular instance.

DR. TEERLINK: I thought you were going to address the rebound ulceration issue.

DR. HIATT: And the absolute blood

pressure issue, as well.

DR. LUND: Rebound, after people are treated, what the relapse rate is? Okay. Well, there is not much data on this, but in Nottingham, we follow people and found a 10 to 12 percent recurrence rate.

DR. TEERLINK: In this study?

DR. LUND: No, no, no, in general.

DR. TEERLINK: So, my point is we don't know what that is. The AGA's consensus document suggests that nitrates increase the rate of rebound ulceration, so that's at least in that part of the literature, and I don't know this literature extensively, but that was what this group had suggested, and we don't have any way to assess that, because that wasn't looked for at all in this database as far as I know. Is that correct?

DR. LUND: No, it wasn't.

DR. GOLDSTEIN: I think it is worthwhile reminding the panel that some of these issues can be dealt with if the drug were to be approved in the labeling and other materials that flow from the

labeling. I think that needs to be kept in mind, as well.

DR. ABEL: If I may, to respond to your question, there is no data available what the recurrence or rebound effect is. The question is does it really make any difference. Quite frankly, again from clinical experience, personal experience, if a patient rebounds or recurs, you have the same option. You can use the available compounded nitroglycerin or you can offer them a surgical procedure.

I still think what is critical is that the--that's not what you are talking about when you are talking about rebound obviously.

DR. TEERLINK: It's recurrence of the ulcer, if they are more likely to have a recurrence of the ulcer when they are treated with nitroglycerin.

DR. ABEL: That has not been my experience at all.

DR. TEERLINK: I have, unfortunately, no experience along this line.

DR. AZARNOFF: The fact is we did follow patients in one of the studies following the end of the trial to see what happened to them, and, in fact, a few patients did develop recurrence, but very few, and there was no difference with whether they had been on placebo or active during the trial.

DR. HIATT: I think there will be a couple of final efficacy questions to clarify. Before we do that--the only question is the blood pressure. Anyone from the sponsor, can you answer what the absolute changes were particularly in systolic blood pressure on drug versus placebo?

Can you remind us when the blood pressures were obtained? I think they were relatively related to dosing, but I can't remember. Just the first one. So, the question is just remind us when the blood pressures were obtained relative to the dosing and secondly, what the actual changes were particularly in systolic.

DR. AZARNOFF: The blood pressures were measured during each visit to the site. We would

have to go back and determine, if we can, when the blood pressure was measured compared to when the subject took the dose. We don't know that offhand.

DR. HIATT: It sounds to me that the hemodynamic effects of this drug, particularly in perhaps more vulnerable populations, older, aren't going to be defined for us today, and in terms of concomitant information, can anyone from the sponsor tell us about adverse events or SAEs in patients who might have been taking phosphodiesterase PDE5 inhibitors?

DR. AZARNOFF: I definitely can tell you that. From Study 1 on, there was a prohibition against PDE5 inhibitors.

DR. HIATT: So, you don't have any data, because there was an exclusion.

DR. AZARNOFF: There was an exclusion and looking at all of the concomitant medication, I don't believe there was any PDE5 inhibitors.

DR. TEMPLE: I am sure it would be contraindicated just like they are now. Can I ask you one thing? The concern expressed was that

there might be some people who really would just go out.

This would obviously be a rare event, but in commenting on the thought that no such reports had come from marketing experience, you said, well, that's not very good, we need controlled trials, but you are not going to have controlled trials on very rare events, so the only source of information, if it were credible information, and I don't know whether it is, really is the marketing experience on something like that, unless you just want to worry about it.

The only kind of data is going to be stuff from post-marketing experience, because it must be a relatively rare event if it occurs.

DR. HIATT: I do agree with that, so if this drug were to cause eosinophilia or something weird, that occurred in 1 in a million people, you would never know that unless you had some post-marketing surveillance.

I think the context of these questions is more along the lines of commonly occurring

cardiovascular things that might occur, the frequency that could be detected, and hypotension in older people may not be uncommon, and there may be higher frequency of that on this drug, or if there was syncope, that might be picked up.

But I realize that given the nature of the population studied, any of those events are so rare that it would take tens of thousands of people to detect any signal. So, for that reason, my recommendation to you all was I think although there are some unanswered questions here, I am not sure how troubling they are.

DR. HARRINGTON: Along the lines that Ron brings up about are there certain populations that would respond adversely to vasoactive effects, women seem to have a higher degree of reported adverse events, the elderly, and I wonder if you have looked to see, is it just the fact that they are women, or when you model it, what women and old people have in common is that they are light body weight, and so are they getting a dose that is more than might be tolerable based on body weight, have

you looked at that?

DR. AZARNOFF: We didn't look at the dose regarding body weight, and you are correct that with some drugs, women do have differences due to body weight, but in pain, there is a large literature indicating that women perceive pain quite differently than men.

DR. KASKEL: Can I say along the lines of gender, any data and safety monitoring information regarding the use in pregnancy? Do we have any concerns? The use in pregnancy, if you had a child-bearing female, would we use this or would it be contraindicated?

DR. AZARNOFF: There was a restriction in all of the trials against pregnancy.

DR. HIATT: Then, I think the other thing I would just like to clarify, there were a couple of lingering efficacy questions. Dr. Teerlink was asking for the absolute benefit, not the relative benefit data in what was it, Study 2?

DR. TEERLINK: Yes, if you could just show the slides from Studies 1, 2, and 3 of the pain

score versus time.

DR. HIATT: Not the percent change, you want the absolute change.

DR. TEERLINK: The mean pain scores through time, which is the primary data I believe that was used for most of the analyses, or at least granted it's mean data.

DR. GIBBONS: All of the analyses were based on the raw data, and as far as I know, I don't have those slides here. We could prepare those slides for you.

DR. TEERLINK: That is what I had asked for before the break. I would have liked to have seen that. I mean it's the primary data of the study that we are basing this decision upon.

DR. GIBBONS: I don't have the data files here to prepare that for you.

DR. TEERLINK: Could you then show Slide 107, I guess.

[Slide.]

DR. GIBBONS: Okay.

DR. TEERLINK: This is the comparison of

the treatments in the three different studies using the same relative measures although we don't know about how the different missing data was handled in the different slides.

Were they all handled the same way for each study in this slide?

DR. GIBBONS: Yes, they were.

DR. TEERLINK: We had mentioned before about the possibility that there was a decreasing treatment effect as we progressed through the study program and as we did more studies, as we had a greater sample size, and I think this does help demonstrate that.

Unfortunately, the reason I was actually wanting to see the absolute number curves, which actually you could get a sense of that the treatment effect was in terms of the Visual Analogue Scale, and we are still not seeing that, and it's confusing to me how you can come to the FDA without information on, or to our committee, without information on the primary endpoint describing that.

So, this is the best I think we can go with it in terms of what the treatment effect looks like as we get through bigger studies.

DR. GIBBONS: You also have seen it in Dr. Hung's report. He had a figure for the third study.

DR. TEERLINK: Study 2, I believe it was.

DR. GIBBONS: Well, he certainly had a picture for Study 3 that had a sea of dots on it with the two lines through 21 days. 104?

[Slide.]

Here, it is. So, for Study 3, you can see what the effect is. The red line and I think that is a black line in the sea of dots will give you the absolute magnitude at any point in time of the differences in the means. That is the smallest differences in the means of any of the three studies, so they would be larger in Study 1 and larger in Study 2.

DR. TEERLINK: And this is the larger study, right?

DR. GIBBONS: This is the final study,

yes.

DR. STOCKBRIDGE: To be clear, that slide was done, that analysis was done with no imputation. That is a line drawn through the mean values on each day.

DR. PICKERING: Could I comment on that? I mean I think what that shows is on day 1, there is nothing below 35, because that was the entry criterion, but that was on that particular day, and the following day, there is a huge spread from zero to 100, so there is a built-in regression to the mean here I think just by the entry criteria.

DR. GIBBONS: It certainly seems that way. I mean it certainly seems that you are seeing the full--that there are a number of subjects who, on the second day, are indicating that they don't have any pain.

DR. HIATT: Do you want to clarify this a little bit further?

DR. LINCOFF: Do you think that represents a variability and that they had to come in with a threshold, and then there is the day-to-day

variability that you pointed out, some days they have a bowel movement, or someone pointed out, some days they have a bowel movement and some days they don't?

DR. GIBBONS: Well, remember that there is a separate indicator for defecation-related pain, and then there is average pain, but certainly average pain is going to be different on the day of defecation. They are supposed to try and differentiate those two. Whether or not they are able to or not is a question. So it may be that some people's average pain is really associated with defecation pain, and then on those days that they don't defecate, they indicate, well, I wasn't in any pain.

So, we have got al of a sudden a bunch of patients who are moving from a 30 or a 40 or a 50 down to a zero.

DR. HARRINGTON: When we were talking about the meaningfulness of the Visual Analogue Scale earlier today, it may have been someone from the sponsor made the comment that you can break

these scales into tertiles, mild, moderate, severe pain.

Is that a true statement?

DR. GIBBONS: I believe the statement was that if we were to do it again, and use a Visual Analog Scale, we probably would have given sort of ordinal markings, if you will, to help guide the user, or that an alternative approach which might be a better choice in the future for these kinds of studies is to just use an ordinal scale with very clearly demarcated labels for each one of the categories.

The advantage of using, you know, strictly from a statistical perspective, the advantage of using an ordinal measurement is that the distance between the boundaries, the thresholds between the categories don't have to be linear, they don't have to be proportional, and so a lot of people shy away from the analysis of qualitative data, or going to non-linear, mixed-effect models, because they feel like there is a reduction of information, but in some cases, it is actually a much more powerful

approach because this idea of a perfect continuous scale is no longer an issue in the analysis.

DR. HARRINGTON: What I am trying to get at, as I look at this scale, I am trying to reconcile two things. The first is the comment from the clinicians who have far more experience than I, that this is a debilitating problem, that these people are really hurting, and so I am trying to put that into context of how they are grading it on this scale.

So, if I accept the earlier statement that the first third of the scale represents mild pain, as I look at this, by day 4 or 5, pretty much the mean score is about a mild pain, and I am not sure I am interpreting that correctly.

DR. GIBBONS: I am not sure there is a hard and fast interpretation for it. There is a lot of inter-individual and intra-individual variability in these pain scores, and those are reflected in the means, as well.

Obviously, we can see that there are still a number of people who are rating their pain on a

90 or 100, I am sure those people are in quite a bit of pain. The means are coming down to the area of around 30 or so. We know that there is going to be a distribution around there.

DR. HARRINGTON: This is Dr. Temple's point, I understand that, but I am just trying to frame in my mind that if we are looking for a modest effect or if we are accepting that there is a modest effect, a modest effect around mild pain, I might be interpreting differently than a modest effect around severe pain, particularly if there is side effect attached to that potential benefit. I am just trying to get my arms around that.

DR. GIBBONS: I think as Dr. Temple pointed out, the shift in the means is invariably followed in a shift to the distribution including its tails. So, it is a shift for people who are currently in severe pain and it is a shift for people who are currently in mild pain, as well, and the shift that we see in the mean is more pronounced in those subjects who are more severely impaired in the initial part of the study.

So, if we look at Dr. Hung's slide here, we see that at 21 days we have a difference of about 3 millimeters. If we look at those people who are more severely impaired to start with, that difference is something on the order of 15 or 16 millimeters.

So, the magnitude of the effect in an absolute sense is very contingent on a variety of different things. The important point is it remains statistically significant even in the aggregate. There are certainly subpopulations for which the effect is much larger than others.

DR. LINCOFF: I was struck actually that I think this distribution is very helpful, if I can point on your slide. If you look at the peak period here, where it seems like there is the most difference, and you look at the patients who are having the most pain, there is markedly fewer in the columns that are the active drug.

There seems to be little difference down below obviously, but there do seem to be a substantially smaller proportion of patients having

the more severe pain, so I think this is an example that just emphasizes how the distribution really does provide more information.

I mean it looks like the sicker patients, there are fewer of them having the really severe pain.

DR. GIBBONS: That is completely consistent with our experience working with these data.

DR. HIATT: Final clarifications on these data?

DR. TEERLINK: This is more an operational point. So, in terms of one of the challenges that certainly this committee has had in the past is when there are symptom measures that are being evaluated by personnel, for example, if the personnel had certain other inputs and things, and then the patients interacting with those personnel, even though it's a person-generated response, their interactions with study personnel who clearly, by your own analysis, seemed to know or have a sense of what's going on with the patients

If we have a minimal statistical effect on an outcome, and we have the specter of some bias being introduced by the study coordinator who is doing the evaluations, how confident can we be that these results aren't really just being generated by the interaction with the study personnel, because they are seen on baseline, day 1, day 14, day 28, and these study coordinator goes through with the patient each time or the investigator does, I don't know. Who is it that actually does the VAS transfer and interacts with the patient in this study?

DR. GIBBONS: You understand those investigators are blind.

DR. TEERLINK: Well, you know, I guess 63 percent of the investigators got it right that the patient was on nitroglycerin ointment as opposed to 42, so there is clearly a differential there, and this is a small differential, as well, and that is the same kind of differential that this committee has had concerns about with other trials where investigator bias can influence subjective scores

by the patient.

I am concerned, and we don't have actually the question being posed to the study coordinator in terms of what they thought the patient was on, so I am just raising that as a concern because there clearly is some partial unblinding here, and that partial unblinding can, as we have seen from other trials, influence patient-reported events.

DR. AZARNOFF: The instructions were to the subjects to complete the diary each night at bedtime. They did that day after day after day. They came back to the clinic for various visits. When they were asked a non-directed question, "Have you had any medical problems since your last visit," it was based on that. If they said they did, they were asked what those were. It would go through a routine diagnosis just like you would do if a patient came to your office.

Those were the requirements for the way the study was carried out. The only other thing that was done is if a subject was not filling out their measurements or not using enough medication,

because it was being weighed, without telling them it was being weighed, they were urged again to follow the amount of drug which they were supposed to use without telling them they hadn't used enough.

DR. FLACK: I would just echo the point where you get unblinded, that despite having a standard protocol, that you can have bias ascertainment from the way people even pose the questions or even prod for the information. I think this is a real weakness, that you are basically dealing with a pretty soft endpoint, and these people did get unblinded, and what this book basically said was the investigators were varied picking it out than the patients were as to what treatment they were on, and I would assume that the study staff may have been somewhere intermediate, but to me it remains a concern.

DR. TEMPLE: I think we have to understand this concern if it's potentially important. What the company is saying is that they are filling out the diary before they see anybody, so that the

investigator or the person in the clinic can't influence that.

Are you reassured by that or not? I want to add one other thing, a question. Do the people filling it out have their old diaries with them, do they know what they said last week?

DR. AZARNOFF: They did not have last weeks with them. They were collected each time.

DR. TEMPLE: I have always felt, I don't know what other people's views are, that you can only cheat if you know what they said last week. If it's just an absolute thing, you don't really know whether 72 is better or worse than what they said last week.

So, one of the things I don't like in PROs is a question, how are you compared to last week, because then if you break the blind, you can cheat or you can be influenced, not cheating, whereas, if it's an absolute score, I don't think you know how to cheat even if you wanted to, or I don't think you know how to be influenced even if anybody wanted you to, because it's just a raw number with

no particular meaning with no--unless they can remember what they said last week.

Now, there was probably a time in my life when I could have, but I certainly couldn't now.

DR. TEERLINK: My point was more that as the study staff knew or were potentially unblinded, that their interactions with the patient would be either more encouraging, more comforting, more like, hey, things are going great, and those kind of things during the physical exam, and those kinds of things, which are the same kind of things that we had concern about with the dyspnea scores and other settings, not necessarily that there would be this general movement, and since this is a fairly soft--

DR. TEMPLE: And they could influence the score even if they didn't really remember what their score was last week, they just sort of feel better about stuff.

DR. TEERLINK: Right, exactly, and that is, to me, the major source of bias in these patients' assessed outcomes.

DR. TEMPLE: There was a big study in the New England Journal about two years ago comparing the results of blinded versus unblinded analysis. Remember that? The thing I am talking about, they looked at about 200 trials in which there was both a placebo and an open treatment, and failed to find any effect of unblinding in symptomatic conditions. I have never been quite sure what to do with that since we still like blinding, but it does suggest the effect isn't huge.

DR. ABEL: John, if I may, as a physician who examined these patients, you know, if I was encouraging I had no idea whether I was encouraging to the patient who was using a placebo or a patient who was using nitroglycerin, so I had really no idea, so if I was encouraging, I was encouraging to both patient populations, and I was disappointed or discouraged, it could have been just was easily of placebo patients, it could have been a nitroglycerin-applying patient, so I had no way of knowing how that would influence patients.

DR. TEERLINK: I appreciate you saying

that, but actually it suggests differently in the questionnaire.

DR. GIBBONS: Let me just put that to rest. We are talking about a difference of, what, 63 percent versus 42 percent or something like that of the investigators, I believe you cited that. There are 17 investigators. That difference is completely consistent with chance expectations in a sample size of 17.

DR. HIATT: We are going to have the ability to debate a lot of this a little bit further, and we have a lot of questions to get through.

The only other efficacy questions I had, just to remind us that the primary endpoint was this pain scale. Do we know if this treatment altered any other kind of outcomes of interest, for example, and we mentioned this, the need for surgery, you know, some other quality of life assessment, some other measurement that would tell us that the change in pain had some other clinically relevant benefit?

DR. AZARNOFF: There were no other measurements along those lines.

DR. HIATT: Okay, so the answer I think was no.

Questions to the Committee

DR. HIATT: Now, we are going to transition I think to the questions unless anyone has any other further things they want to clarify with the sponsor.

In this section, the committee usually discusses things and would call upon you if there are things to be clarified and if you feel we grossly misrepresented something, please let us know.

At this stage, we actually would project these questions and have a discussion. There are three voting questions, and then there is a lot of questions around opinions.

Forgive me for this one little digression, but there are a lot of new people on the committee, and my experience has been just modestly a bit more than yours, so I am not trying to tell you what to

do here, but I think what our charge in this kind of environment is frankly to look at the evidence specifically and judge the evidence for what you are seeing.

Some comments have been made today that maybe there is a lot of off-label use of this, there may be a great unmet need, there may be a lot of other things going on in clinical practice that we all care very much about, but our goal today is to judge the evidence, and that is really what we are supposed to really adjudicate. So, I just wanted to make that statement.

There is as bit of a long preamble here. I can read this if anybody would like me to, but you can all read this yourself.

Dr. Koltun, are you on?

DR. KOLTUN: Here.

DR. HIATT: Can you read silently while I read aloud? Do you see the questions? The first slide is Questions to the Committee.

The committee has been asked to opine on Cellegesic. Do you see that?

DR. KOLTUN: I can't really read it, no.

DR. HIATT: Do you have a copy at your desk?

DR. KOLTUN: Yes. What page exactly is it?

DR. HIATT: It's page 1 of the questions. We want to make sure you are on the same page. So, let us know if you have any ambiguity about that.

DR. KOLTUN: Yes, I have got it.

DR. HIATT: The first two paragraphs are really a bit of a background on some of the evidence we have looked at. The next paragraph, there are two issues for the committee to address. At the end, we are asked to choose among three outcomes. I want to just emphasize what these three are: Approval, Approvable, and Not Approvable. Just to distinguish approval from approvable is shown here.

Any questions about that?

A lot of drugs get approvables, and they don't go to market.

Bob, do you want to explain that?

DR. TEMPLE: Actually, the law has been changed to urge us to just give approved versus not yet approved and here is why, but we haven't finished changing our rules that way.

So Approval means you are ready to go as soon as you can market.

Approvable means that there is something missing. It could be a lot of different things. There is no real rule about it. It could mean you need to do another study. It could mean you need to do more analyses to reassure us that this was not a problem.

As a general matter, it usually means there is what we would certain call some evidence, but maybe not quite enough.

Not Approvable means there is a horrible safety problem that really we don't think you can overcome or that there is just no evidence at all.

DR. HIATT: There is futility.

DR. TEMPLE: Yeah, but the gap between them is variable enough, so I wouldn't want to say what it always means.

DR. HIATT: Okay. So, with that as a background, the last bit of procedure here is I would like to kind of reverse orders around the room. I don't want everyone to go in the same order and bias other people's thinking.

I was actually encouraged once to just have people write this down on a piece of paper and then declare what their vote was later, but that's too hard. I think just try to be cognizant of some level of independence when you discuss this, and to remind you there are three voting questions.

Let's begin with the first question.

The sponsor believes Study 2 should have been considered persuasive, because the post-hoc inclusion of a quadratic term in the regression analysis was justified.

What is the interpretation of the linear term in an analysis with a quadratic model? That is the first question. It is not a voting question.

Dr. Goldstein, you have the ability to discuss, but not vote. So, we can start with you.

DR. GOLDSTEIN: Thank you, Mr. Chairman. As industry liaison representative, I am hardly one to be considered a disinterested party at these proceedings, but the only comment I would make at the outset is something Dr. Harrington said earlier in the meeting resonates, and that is before I was in the industry for 30 years, I practiced for 17, and the relief of pain, particularly acute pain, when there were few other options available, and those options had difficulties in and of themselves, as has been clearly delineated here today, made me feel, shall I say, in some sympathy with Dr. Harrington, that the clinical mattered a lot more than the statistics.

While I have a healthy respect for both, I have to lean toward the clinical, because those are my roots, and I would certainly recommend that the panel remember that we are in the business of treating human beings, large and small, who are in pain, and who if they choose to elect surgery, will run a 35 percent risk of coming out of it having difficulties with bowel control, which if you have

ever had it, and I have a daughter with that for other reasons, is not a pleasant thing to contemplate.

So, keep in mind we are treating human beings in acute pain. Thank you.

DR. HIATT: Thank you.

John.

DR. FLACK: All I know is that in the study in question about the interpretation of linear term, up until the last six days, it looked linear, and I will trust my eyes and let the more sophisticated statisticians try to tell me what it means or exactly what that term means.

I think that they did the right thing in putting the quadratic term in there, so I think that was fine, and it wasn't cheating, wasn't anything malicious, and all, and gave the best analysis of the data even though they couldn't predict it upfront.

DR. HIATT: David.

DR. DeMETS: Well, as has been outlined before, the linear term in that quadratic model

doesn't translate into rate of change. Rate of change is something different. So, while I would agree that it was a sensible thing to do to put the quadratic term in, you can't just look at the coefficient of the linear term in that model and equate that to rate of change. It is more complicated than that.

DR. PICKERING: To my naive way of looking at them, they all looked non-linear to me, so I guess I would generally favor the quadratic model. I mean I guess if there was a big effect, it probably wouldn't make much difference which model you used, and a lot of what we have heard today is you get different analysis, different conclusions depending on which type of analysis you use, which to me means there is not much of an effect there to begin with.

DR. PORTMAN: I don't really have a whole lot to add on that.

DR. KASKEL: I think anything that obviates the need for surgery should be attempted even if it's a short term gain, at 21 days, and

reevaluated.

DR. FINDLAY: This is way beyond my area of expertise to discuss the statistical issues, but it seemed like from the discussion, that it was a reasonable methodology.

DR. WARNER-STEVENSON: As I understand the answers to the questions this morning if one used a linear model for the first 21 days, it, in fact, would have looked like there is a significant difference, and I think it is hard to know how long it takes things to get better.

Einstein said, "If we knew what we were doing, it wouldn't be research, would it?"

DR. HIATT: I can just echo that. I think in the study where the quadratic model was used was the best fit for the study, but in terms of the totality of the evidence, I am not sure what the best fit would be, and simpler would be better. If it was a linear fit, it should be relatively robust.

So, I think what the data are probably telling us is that the response to the drug differs

from early to late, and that is how I would interpret that. Whether that should have been in any interpretation of a linear response in this context, I think would be challenging.

DR. TEERLINK: I think Bill just hit upon the point that I was going to make, and that obviously, the effect does seem to change through time, if, in fact, there is effect. I think we want to be careful in giving people the future license to pick the model that fits your data best, or if you are going to let us do that, that's good, because then I will be happy to do it, but I don't think so.

So, obviously, we need to be careful of that, but I think in this specific case, it is an appropriate approach to the data as it evolved.

DR. HARRINGTON: Like John, I will defer to the statisticians in the room with more expertise, but I was comforted by David's comments that you should choose the best model that fits the data with some caveats around that, as John I think is indicating, that we don't want people to be

taking liberties with what they find after the fact.

I do agree with Bill and others that what I can determine in looking at the data, that the timing of the effect does appear to differ. I am a big confused still by the fact that the data looks one way in Study 1, a different way in Study 2, and perhaps somewhere in between in Study 3, which leaves me a bit on the margin with a lot of studies that are on the margin.

DR. LINCOFF: Although I remain concerned about the process of changing endpoints of the question, I do defer to Dr. DeMets and Dr. Flack, what I believe is an important distinction that making the best mathematical fit may not violate that principle.

That being the case, I think that clearly these curves fit a quadratic equation better than a linear equation, and it seems the interpretation of this is that there is more effect early on, but I think that is fine clinically.

There is certainly clinical benefit in

making people's pain go away sooner even if it's ultimately going to go away at the rate. As we say, all survival curves ultimate merge, and the same thing probably for pain, so I think in terms of clinical relevance, it certainly is relevant, and I guess the best interpretation of this is that it is an early effect, not a late effect, and that's the best fit of the existing data.

As for the different studies, I think that Study 1 was just very small. I mean you can look at those curves and almost put anything in a shape into it. Study 2 and 3 both are curvilinear, so I think that that's probably the real data.

DR. TEMPLE: I just wanted to ask David for a clarification. You thought introducing the quadratic term was reasonable because it was called for by the non-linear nature of the data, but then said you weren't quite sure what the linear term--oh, sorry--and that that is therefore a legitimate way to conclude that there was a difference between the treatments? Does that follow or not?

DR. DeMETS: I didn't say that second part.

DR. TEMPLE: Oh, then, in what sense is it appropriate, because that is what they concluded from that analysis, that the treatments were, in fact, different?

DR. DeMETS: That's 1.2. I was talking about 1.1, right?

DR. TEMPLE: Okay. When you said it was appropriate--

DR. DeMETS: You should represent the data as best you can. I mean that is what we train statisticians to do, to do data reduction in the best way we can. In this case, the quadratic term changes the interpretation of the coefficient of the linear term. It is not rate of change anymore. As Jim outlined in his report, the rate of change is the derivative of the function.

That is a slightly different question. I know it's subtle, but it's a different question.

DR. TEMPLE: So, you are saying you can't just look at the linear term the same way you

ordinarily would.

DR. DeMETS: The strict interpretation of the coefficient of the linear term in that quadratic model is in the presence of a quadratic term, this is the linear contribution.

DR. TEMPLE: All right. But as I was trying to ask Jim before, why is that so important?

DR. DeMETS: The question was posed as rate of change. That was the question that was posed as the hypothesis for the study, perhaps not as well thought out, that we all would like, but that is the way it was posed.

If you have a linear model, then, it's a slope. That's easy.

DR. TEMPLE: So, if you used a different model, then, it's not a slope anymore, but it still reflects some comparison of the two curves.

DR. DeMETS: Sure, but it's just not a simple interpretation. I mean a different way of asking the question, do these two dose-response--not dose response--two curves change would have been to do an overall likelihood ratio

test or something of that nature, that these two curves are different. That is not what was done, but that would be a natural thing to think about had you known the data were going to come out this way.

DR. TEMPLE: So, when they introduced the quadratic term, that alone didn't do what you are talking about.

DR. DeMETS: No.

DR. TEMPLE: When you introduce the quadratic term and conclude that the two curves are different, what exactly are you concluding then, because that's what they said, you know, p less than 0.05.

DR. DeMETS: I suspect--

DR. TEMPLE: It means they are different.

DR. DeMETS: I can't point to the page right now, but I think it's true that the coefficient, the linear coefficient and the quadratic coefficient are different. Am I remembering this incorrectly, Jim? So, I think that they said that the two coefficients are

different, therefore, the curves must be different, which is true.

DR. TEMPLE: And what conclusion does that lead to, or is that 1.2? I understand the point that you can't just look at the linear terms and say, oh, this is 30 percent better, but you have got to do something.

DR. DeMETS: Neither Jim Hung's analysis, nor the sponsor's analysis, at least as far as I remember, gave us an analysis are these two curves different.

DR. HIATT: Can't we say, Bob, that these curves are best described as the rate constants are changing over time, and that the curves are different, but you can pick one point on the curve and extrapolate that to a linear regression, but the point is it is not linear, it's really a quadratic, therefore, its rate constant changes over time.

DR. TEMPLE: Right, but in a certain sense, we don't really care. We just want to know if one of them is changing faster than the other.

I mean that's all we care about is the pain getting better faster on treatment than it is on non-treatment. I mean that's why you are doing this.

DR. DeMETS: Then, take the derivative and compare those, but you are asking is the rate change different.

DR. TEMPLE: No, what I am really asking is if it was appropriate to use the quadratic term and do an analysis, what have you learned from that analysis? It gets a p-value assigned to it, so somebody thinks it tells you something.

DR. DeMETS: There is now multiple p-values. There is a p-value for the linear term, there is a p-value for the quadratic term, there is a p-value for the likelihood ratio, which you are looking at the overall curve. You can either do, I suppose, take a derivative of the function and compare those two derivatives, the slope of this curve, and get a p-value for that, so now there is four p-values. There is lots of p-values is the point.

It goes from a very simple question to a non-simple question very quickly with that additional term, but yes, you would do this naturally to summarize the data. It would be a sensible thing to do, there is no question about that.

DR. TEMPLE: But there would be no conclusion you could reach from it.

DR. DeMETS: By itself, you don't have a conclusion. You have to go further, what do you mean about the curve.

DR. GIBBONS: Just to maybe bring it back, one conclusion is, is that you have at least an admixture of two rates, maybe an admixture of an infinite number of rates. When you look at the first 21 days, however, the linear component alone without the quadratic term differentiates the drug and placebo.

So, at least we know that through 21 days, the function is linear, and the difference in terms of rate of change is statistically significant, and I think that is a good--

DR. TEMPLE: But you didn't need the quadratic term to reach that conclusion. You could have just looked at the first 21 days.

DR. DeMETS: That is correct, but that was post-reflection.

DR. HIATT: Yes, that is all post hoc.

DR. TEMPLE: I know it's post hoc, but some people have said that the analysis including the quadratic is so logical in terms of the data that it doesn't need to be dismissed as post hoc, which is what I understood you to be saying, David.

DR. DeMETS: That is true, but it doesn't answer your simple question, at least as far as I can understand it.

DR. TEMPLE: Oh, that's okay. What question does it answer?

DR. DeMETS: You are seeing two curves and you are asking are they the same, but no one did that analysis.

DR. GIBBONS: We did do that analysis and it was reported. We tested to see whether or not those two curves were the same by including both

linear and quadratic effects, and both the linear and the quadratic effects were statistically significant. That is the treatment by linear and the treatment by quadratic effects were both statistically significant.

DR. STOCKBRIDGE: Why stop with polynomial with three free parameters? Why not go to a 20-degree polynomial? I mean the curves are clearly different, the points don't overlay one another. If I add enough frequency parameters, I can separate them.

What is it about picking the number of free parameters in the model after the data are in-house that lets you tell me with confidence that these are different?

DR. DeMETS: You are asking me?

DR. STOCKBRIDGE: Sure.

DR. DeMETS: Well, we again teach curve fitting to our students, and you ask how many parameters you need, that you don't get any reduction in the variability anymore, and when you don't have any more reduction, you stop. That's

how you curve fit.

DR. STOCKBRIDGE: Okay. But if the number of free parameters is equal to the number of data points, I will end up with an exact fit.

DR. DeMETS: Of course, but I presume your data-fitting algorithm would stop way before then.

DR. GIBBONS: And even if you did use a 20-degree polynomial, in fact, it is very common practice in mixed-effect models to not assume any curve whatsoever, and just simply use simply contrasts of each point relative to baseline, which gives you a non-parametric comparison of the two curves.

It is the treatment by time interaction that's important, and if you really think you need 20 degrees of freedom, you are going to pay a price in statistical power to be able to find a significant treatment by time interaction if you are loading up either with simple contrast or polynomials. That is a very dangerous thing to do. It will tend to be more conservative.

DR. TEMPLE: David, what I was trying to

elicit was because, you know, what can you learn from Study 2 about whether the treatment is better than placebo. What I first heard you saying was that even though it was after the fact, and not anticipated, it was reasonable to include an analysis with a quadratic model, the conclusion of which presented to us is that oh, the curves are different when you do a quadratic model, different in some way without worrying about what that means, which I guess I thought translates to oh, these is a difference between the treatment and the placebo.

If you then believe that, you might be willing to look at the first 21 days when it looks linear and draw a conclusion from that, but I guess the question is what is one able to do without violating the usual proscription against diddling the data after it's in your hand.

You can just answer yes or no.

[Laughter.]

DR. DeMETS: Well, I think we are engaged in some level of diddling, but nevertheless, what do I learn? It was a non-linear response over the

long time period, and it appears that the first 21 days is linear, and perhaps the rest is linear, there is at least two or perhaps more, and that gives you a very good clue as to what you would do next, which is what was tried.

So, you learn a lot from Study 2, but by itself, it has its complications unfortunately.

DR. TEMPLE: One of the questions obviously, it is not put that way is if you were to believe that Study 3 won and is informative, how much does Study 2 add to it. I mean if it generates the design of Study 3, that is a very virtuous thing and everything, but that doesn't represent a contribution to the overall data, or maybe it does in a smaller way. I mean that is sort of what this is really all about. You would say the same thing about Study 1, I suppose, too.

DR. HIATT: Are we comfortable to go on to the voting question? Oh, Dr. Koltun, sorry. Dr. Koltun, do you want to weigh in on 1.1?

[No response.]

DR. HIATT: Do we have clarity on the

issue? I think these data over their totality, over the duration of the extent of follow-up don't look linear to me, and therefore, as I think Dr. DeMets is saying, you know, the best fit is the best fit, I mean the best representation of the data.

The thing I am struggling with is prospectively, defining whether that was really exploratory or confirmatory.

If we are ready to go, let's do 1.2. Was the quadratic model the proper analysis for the purpose of decision-making?

We will start at this end. Do you all want to comment, or do you want to let us go ahead and vote? Okay.

DR. LINCOFF: Pretty much what I had said before. I believe that within those constraints, it is the proper--yes is my vote.

DR. HIATT: So, just so I understand that, when you say "yes," then, that is a positive pivotal trial?

DR. LINCOFF: No, the question here, I

mean if you want me to try to answer that, that is not--

DR. HIATT: Okay. So, your answer is a simple yes?

DR. LINCOFF: Yes.

DR. HARRINGTON: Maybe because seeing where you are going with that, Bill, maybe Norm can describe for me what do you mean by this question, for the purpose of what decision?

DR. STOCKBRIDGE: Well, if you think you need to identify more than one study that won on its primary endpoint, you have got to figure out what the prespecified--well, maybe you don't care what the prespecified analysis was for Study No. 2.

That is sort of what I am hearing is that it's okay to look at the shape of the curves and figure out what the right analysis is to do of a study, but somehow or other you have to got to come to a conclusion that there is a p-value you can assign to study No. 2, to decide whether it is positive or negative.

So, I am asking whether the quadratic

model determined post hoc was the right way to figure out whether Study No. 2 was positive or not.

DR. HARRINGTON: That is exactly the context I was trying to emphasize.

DR. LINCOFF: I still maintain a yes vote.

DR. HARRINGTON: I think that as I look at Study 2, well, as I look at the three studies, I see them as a progression based on exploration, and that really 2 was informative to help design Study 3, but I wouldn't accept 2 as a definitive trial.

DR. HIATT: So, your vote on the question--

DR. HARRINGTON: Would be no.

DR. HIATT: --would be no, okay.

DR. TEERLINK: I guess put in that context, my answer would be no, as well, though a qualified no order in terms of I think it does--given that we still haven't talked about treatment effects, as well, and things, it would be a qualified no.

DR. HIATT: Dr. Koltun, are you back on?

DR. KOLTUN: Yes, I am.

DR. HIATT: We didn't get your impression of 1.1, and we need your vote for 1.2.

DR. KOLTUN: I am not going to pretend to be very sophisticated in my biostatistical analysis, but my impression is that the quadratic analysis had enough arguments to justify its use.

DR. HIATT: So, that is 1.1. And your vote on 1.2?

DR. KOLTUN: Is 1.2--my book says 2.

DR. HIATT: Maybe you don't have the latest. 1.2 is bolded and it says, "Was the quadratic model the proper analysis for the purpose of decision-making? Please vote."

DR. KOLTUN: It's the same vote as 1.1.

DR. HIATT: Which is yes?

DR. KOLTUN: Yes.

Can I make one comment?

DR. HIATT: Please.

DR. KOLTUN: I believe the industry rep, I think Dr. Goldstein made some comment about the risk of surgery being 35 percent. That's not the actual number. The statement it being up to 35

percent is correct, but the actual risk of surgery and incontinence, which is usually spotting, is about 1 percent for the surgery, although it makes no difference to this question, but it was raised. Dr. Goldstein talked about it in this question, so I thought I would just make that comment.

DR. HIATT: Thank you.

My vote on this question is no. I believe that there is a suggestion of a signal in the totality of the data, but I am concerned that if the models have to change every subtly to understand what that signal is, I get a little concerned. I think this decision was post hoc, once again trying to just stick with the data in the intended development program, I would vote no for 1.2.

DR. WARNER-STEVENSON: I would vote yes. I think it's unfortunate that the protocol analysis didn't pre-specify that they would use what other method fit the data, but I still think we should use the method that fits the data. So, it is yes.

DR. FINDLAY: It strikes me from this

discussion that it clearly wasn't regulatorily proper, but it strikes me that we decided that it was sort of acceptable, or at least I have come to that conclusion. I would vote yes.

DR. KASKEL: It was post hoc, but I would vote yes.

DR. PORTMAN: I would vote no. I am a clinician, and it worries me that when I look at that curve, there seems to be two different effects, and it has been brought up before that it looks linear to 21 days, and then it doesn't look linear, or it takes off in a different point, and I think that may be a biological difference.

So, it bothers me a little that we changed the statistical method to describe that, and suggest that it works for the whole 56 days. I would buy that it works for 21 days, but I don't buy that it works necessarily for the whole 56.

So, while I would be willing to say this is an acceptable trial, I would limit that to the first 21 days.

DR. PICKERING: I would say yes, it was

the proper analysis even though it wasn't the prespecified one.

DR. DeMETS: I think that it was probably a reasonable analysis, but because of the nature of how it arose, I would say by itself it was not sufficient, so I would vote no.

DR. FLACK: That is helpful after all the discussion we just had with you.

[Laughter.]

DR. FLACK: Those curves are different for most of the time, they were different at 21 days, and models are one thing, but to me, most of the time these curves are different. I think it is appropriate, and so I would vote yes, I don't think it's unreasonable.

DR. HIATT: Please comment.

DR. GOLDSTEIN: Actually, I was going to comment for the record what I thought I said on Slide 19 on page 10. It says, "Postoperative incontinence up to 35 percent."

Now, that is the comment I meant to make. Perhaps it came out garbled, but I was referring to

that slide.

DR. HIATT: John, we just missed your vote. For clarity, your vote is?

DR. FLACK: Yes.

DR. TEMPLE: I just thought it seems worth observing, we have no data at all on whether this drug keeps anybody from going to surgery. It would be nice to know, but we don't have anything like that.

DR. HIATT: Let's move on to Question No. 2. This is a discussion question. Study 3 called for a Last Observation Carried Forward analysis of pain data from subjects who discontinued "due to headache."

The sponsor interpreted this to mean treatment-related headache, leading to the previously cited p equals 0.0498. Various alternative analyses are summarized below from Dr. Hung's review of July of 2004.

There are six conditions: LOCF for withdrawal for drug-related headache, as stated 0.0498; add all data available for 1 subject, that

p value goes 0.0843; LOCF for withdrawal for any headache, 0.12; LOCF for any withdrawal, 0.0943 to 0.15; no imputation, 0.0489; no imputation and no post-withdrawal data, 0.0309. You have seen all of that.

Is the analysis based on "drug-related" headache a reasonable interpretation of the protocol? Is it reasonable to expect that the determination of drug relatedness would be unambiguous? I think we will just ask or comments here.

Dr. Goldstein.

DR. GOLDSTEIN: No comment.

DR. FLACK: I think the definition of drug-related headache is very arbitrary, highly accurate, and my answer is no.

DR. DeMETS: I would echo the same sentiments. It is never easy to have cause-specific anything, it's interpretation, and I would prefer to take all withdrawals as post-headache withdrawals.

DR. PICKERING: I would agree with that,

yes.

DR. PORTMAN: So would I.

DR. KASKEL: Ditto.

DR. FINDLAY: Same.

DR. WARNER-STEVENSON: No.

DR. HIATT: So far we are all consistent. My interpretation of drug relatedness during the blinded phase of the study is hard to interpret, but it is easy to figure out once you unblind. So, I think it should be any headache.

Dr. Koltun.

DR. KOLTUN: I would agree. Nothing can be unambiguous.

DR. TEERLINK: No, and it should be based on all headache.

DR. KOLTUN: Pretty much.

DR. HARRINGTON: I agree with the previous remarks.

DR. LINCOFF: I agree although it may be a null point if we later believe that the LOCF wasn't the best analysis, but no, I agree.

DR. HIATT: Comments from the end of the

table? Any comments on that? No?

DR. TEMPLE: No, but I think we agree the bigger question is was LOCF the right analysis even though they, under some duress, agreed to do it.

DR. HIATT: Maybe we should go to 2.2. The sponsor's backgrounder comments extensively on the use of LOCF with a mixed-effects model. Should LOCF have been included in the analysis?

We will start back on this side.

DR. LINCOFF: LOCF seems to me to be an attempt to maintain as much of an intention-to-treat, that is, to include data on patients that leave, so that you can compensate for the reasons they might have left.

I think for symptomatic treatment, such as this, particularly if we don't believe that there are substantial safety issues that could be contributing to leaving that would have a huge impact, and I realize that headache may be a safety issue, but nevertheless, I think on the balance, I would rather know if there is a signal under therapy that is not confounded by extrapolating out

with an LOCF.

So, I believe it would have been better not to have used the LOCF.

DR. HIATT: So, just to clarify that statement, so if you believe that it is better not to use it, you actually think the mixed model?

DR. LINCOFF: I think the mixed model without the LOCF would have--

DR. HIATT: Was the best model?

DR. LINCOFF: Yes, was the best model for the information we would like to have for this. So, that would be a no as an answer to this then.

DR. HIATT: Fine.

Dr. Harrington.

DR. HARRINGTON: Again, I will defer to some of the statisticians in the group with more expertise, but I thought that Dr. Gibbons made a pretty compelling argument that the mixed-effects model has other methods of including all of the data that are available that would not require inclusion of the LOCF, so I would agree with Mike that I do not think that they should have been

compelled to include it.

DR. HIATT: John.

DR. TEERLINK: I think that there is good arguments actually on both sides to take kind of the Buddhist middle way here, but given that the LOCF can also account for the patient having these headaches for other reasons, to stop and carry those symptom aspects that make them quit, as well, that that would be an aspect that is accounted for by the LOCF.

So, I would actually be fine with the mixed-effects model, I am fine with an LOCF. The fact of the matter is that the primary endpoint was predefined to be the LOCF, and that's the way I think the data should be interpreted.

DR. KOLTUN: I am relatively indifferent on this point.

DR. HIATT: I can understand. I think historically, LOCF is the proper analysis for these kinds of studies, and in particular, we are a little bit worried about potential unblinding and where headaches may differentially affect things,

but on the other hand, I think the mixed models is a better representation of the data, and something I would like to look forward to drug development starting to use.

I actually think that in this particular case, I will agree with the sponsor's use of mixed model.

DR. TEMPLE: I was just going to ask. Maybe everybody is thinking of this as it goes by. LOCF, as they planned it, involved adjudication of whether the headaches were drug related or not. We were all very suspicious of that designation, but that was, in fact, the plan for better or worse.

As you can see from the table, if you drop everybody and do LOCF, they don't make it by nominal significance. If you do it the way they planned it, they do, and if you do mixed-effects modeling, they do. So, it gives you a lot of choices.

But I assume when people say that they either do or don't like LOCF, that was as planned or as improved by dropping everybody with a

headache.

DR. HIATT: I think as planned, the LOCF is the right answer, but I think in terms of the way they approached these, the fact they have got multiple data points across the time interval, I think they are using all the data, I think that's a good thing.

DR. TEMPLE: I understand that, but for those people who think that LOCF was reasonable, it's important for us to know whether they mean LOCF as done with this somewhat dubious description of whether the headache was, in fact, drug induced or not, because only three of them were, and all the rest of the headaches, which were more numerous in the treated group, and which weren't officially designated as headaches for which you did LOCF, that gives a different outcome. I am just taking note of that.

DR. HIATT: I think the committee felt pretty strongly that when you look at that issue, it should be all headache, not drug related, which gives you the 0.12 p value for LOCF.

DR. TEMPLE: Okay. But as people go by, I think they should say which of those two things they mean.

DR. WARNER-STEVENSON: Can I abstain? I don't understand the difference well enough between mixed-effects model and LOCF.

DR. FINDLAY: I will abstain, too.

DR. KASKEL: I think I would like to abstain.

DR. PORTMAN: You can't do what I was going to do. This is a tough statistical question obviously, that it is difficult, you know if you don't have that expertise. It is a critical point as Bob just made, and I don't have the expertise, so I will abstain.

DR. HIATT: I just caution us that I think, though, it is a statistical construct. I mean the reality is, and as I tried to say in my introduction here, that headaches were differential in terms of cause a dropout. We all agree that you can't distinguish one headache from another, and so a more conservative way to deal with that I think

would be the LOCF for all headache, and that makes it a negative trial.

I think that's the context. We can argue the statistical methodology at some other time.

DR. PORTMAN: Well, that I agree with. I mean it's just the phrasing of the question that is tough for me to answer.

DR. HIATT: Let's try to stick with the intent.

DR. PICKERING: As I understand it, the actual observation that was being carried forward was the anal pain, previous anal pain observation, and what bothers me about using the LOCF is that we have seen that within one day, somebody's pain score can go from nearly 100 to nearly zero, which I think is unusual. If you are looking at blood pressure data or something, you don't see that, but in this particular case, it may have introduced a lot of noise, so I guess I would vote no even though that was the agreed-on primary technique for analyzing the data.

DR. HIATT: I think if I got this right,

the people that dropped for headache tended to have a bit more anal pain, am I right about that? So, if you carried that kind of observation forward, you are going to hurt your drug difference essentially, having to explain why the p value got to 0.12. I think I got that right.

Dr. DeMets.

DR. DeMETS: I can't abstain? Well, I don't know exactly how the LOCF got into the protocol, into the plan, but generally speaking, when you do that analysis, which I think is what Dr. Hung did is a sensitivity analysis, you are trying to find out if the missing data matters at all, because most statisticians, I think are suspicious of missing data especially due to withdrawal and whether it is really independent or not.

So, I am not a big fan of LOCF, and I think that the mixed-effects model is one that I said many of us have used for a long, long time, but the problem is does any of that missing data matter, and the LOCF says it could.

You know, we don't know what the right answer is, but it could matter. So, for me analysis No. 2 is the one that I like, because I want to use all the data that I have available, I wish I had more. So, how does that No. 2 come out? Well, there is at least one subject I guess that had data post-withdrawal, and it was used.

But that is exactly what I was appealing for earlier. I want complete follow-up on all patients. Off treatment does not mean off study. So, this list of analyses, what I do if you handed this data to me today, is I would probably do analysis similar to No. 2 and wish I had more data.

DR. TEMPLE: Can I ask about that? This is analyzing in a symptomatic condition, analyzing the person who has been removed from the study in a symptomatic condition.

DR. DeMETS: Yes.

DR. TEMPLE: I mean you know that is just never done.

DR. DeMETS: I understand that.

DR. TEMPLE: Okay.

DR. DeMETS: But then we have these discussions.

DR. TEMPLE: Well, you know, you do that in a blood pressure trial, I will tell you what you are going to get. You know, you will just get a bias toward the null, because as soon as you take the blood pressure drug away, the blood pressure is going to go up. Is that really useful?

DR. DeMETS: You have all kinds of bias.

DR. TEERLINK: What is the intention to treat?

DR. TEMPLE: This is intent to treat, which we advertise all the time for outcome studies, but never, with one exception that I could mention, impose in symptomatic conditions because the assumption is you take the drug away, the symptom comes back. You take the blood pressure drug away, blood pressure comes back. What have you really learned?

I guess it is appropriate punishment for not getting people to the end of the study, but it doesn't seem very informative, it doesn't give you

a good estimate of what the true effect size is, and I must say this is a--

DR. STOCKBRIDGE: ITT never does that.

DR. TEMPLE: I know. You pay the price in an outcome study because the events are so important. That is why you do it. You know it's a bias toward the null, but you do it because it is so important. In symptomatic conditions, I mean I just don't agree with that, but that is a longer discussion.

DR. DeMETS: The one other--I guess I shouldn't interject here--but as was pointed out, there are a lot of points in these regressions, right? It is not just two or three points. There is 20 or 30 points in there, so the point that is missing, no matter how variable it is, perhaps has less influence on the analysis than the whole.

If you take something as extreme as LOCF, you are going to have an impact, because taking that one point, moving it all the way over to the end, and then fitting a line through it, its most influential point probably in the whole data set.

DR. TEMPLE: Well, I am not advertising LOCF either. I think in a case where you are doing multiple measurements, that doesn't make sense, but No. 2, you are adding the available data for one subject, you have seven or eight more who dropped out early, you don't have their data, so they are not going in.

But do you have a view about the no imputation approaches, or I guess you could say you want those, but you want the one patient with data added into that. I don't know that that has been done.

DR. AZARNOFF: It hasn't.

DR. HIATT: John.

DR. FLACK: The way I learned how to do clinical trials is that you get the data on people even if they drop out of the study, but I also know the practicality is in the field. Once people get off drug, there is very little emphasis on bringing them back. The coordinators don't do it, and many times even the monitors don't really push for it, so what you end up with is here, you end up with

really no data points on those people.

You get them rarely like we have here, but on the other hand, if I am going to treat 100 people with something, and I have got 30 dropping out, I am not sure it's just as informative to look at 70, whether it's symptomatic or not and say what happens. I think that is secondary analysis, but practically, we just don't end up with that data because people don't get it.

Not to belabor the point, I would just vote no on the LOCF.

DR. HIATT: Thank you.

Comment?

DR. GOLDSTEIN: I have joined the loyal company of non-statistician abstainers.

DR. HIATT: All right.

The last discussion point here. Subjects enrolled with one kind of pain and discontinued with a different pain. Was LOCF conservative enough?

I didn't do 2.3. Sorry.

A few subjects had data following

discontinuation. Should this post-discontinuation data have been included in the primary analysis? We were sort of touching on that just now, so we will start back on this side of the room.

Dr. Goldstein, do you want to make a comment on that?

DR. GOLDSTEIN: No, thank you.

DR. FLACK: Not to belabor the point, yes.

DR. DeMETS: Yes.

DR. PICKERING: Yes.

DR. PORTMAN: Yes.

DR. KASKEL: Yes.

DR. FINDLAY: Yes.

DR. HIATT: Yes.

Dr. Koltun.

DR. KOLTUN: Yes, intent-to-treat should be yes.

DR. HIATT: Yes is your vote?

DR. KOLTUN: Yes.

DR. TEERLINK: Yes.

DR. HARRINGTON: Yes.

DR. LINCOFF: No. I am sorry, I hate to

be the new guy, but our question here is if the patients are on the therapy, do they get better. It is not like a therapy where they may die as a side effect.

If they can handle the headache, will they get better, and I think including data once they are off the therapy, where there could be no expectation that they would continue to have pain relief, it is not helpful, so I must say no.

DR. TEMPLE: I just have to point out it is simply never done. All the symptomatic treatments that have ever been approved, I only know one Alzheimer's study where because Paul Meier was on our case, we had to follow the patients out, but it really is not done, and I guess my reaction is there are many sources of error in trials.

This celebrates the one source of error that might happen because of informative censoring or informative dropouts, and ignores the gross underestimate of the effect size that you are going to have if you include a bunch of people that aren't getting the drug anymore.

I think it needs more discussion at a minimum, and it is a total change from the way all trials have been done.

DR. HIATT: I will change my vote.

2.4. Subjects enrolled with one kind of pain and discontinued with a different pain. Was LOCF conservative enough? We will start back on this side of the room.

DR. LINCOFF: I don't really understand the question. I am not sure what LOCF has to do with that. I mean are we being asked should we have included a global pain score here? I am not sure what the two halves of the question have to do with each other.

DR. STOCKBRIDGE: But that is the fundamental problem, isn't it, that there was no global pain score, so Bob, sitting over here, really wants to know what the effect of the drug was on pain, so he really doesn't like that post-treatment data are included or LOCF was included.

He is interested in that aspect of this.

I think there is a real problem in any analysis that pays no attention to the fact that the withdrawals in the first 21 days were from drug. They were presumably not people who got completely better and left the trial.

They were people who left early because they didn't tolerate the treatment very well, and somehow or other, this is your chance to get some kind of net benefit assessment is by including them and in some kind of analysis where you get penalized for leaving the trial early. That is really what the effect of LOCF was, and the question is whether that was enough.

DR. HARRINGTON: So, Norman, are you asking, given the choice, would you prefer analysis 1 versus 3 or 4, is that essentially what your question is?

DR. STOCKBRIDGE: Well, no. You could have said that people who left the trial because they couldn't tolerate the treatment should have gotten something worse happen to them in the analysis. They could have gotten worst rank. They

could have gotten, you know, some bad score assigned to them as a way of penalizing you for not being able to tolerate the drug.

DR. HARRINGTON: Is there a standard methodology for that?

DR. TEMPLE: No, we usually--but this is a little different because they are both pain. We usually discourage analyses that try to mix good things and bad things because, among other reasons, they happen in different patients and people put different values on them, and my view has long been that you should try to separate those two things, and not throw them together.

Can I make just previous comment? The alternative proposed by the sponsor, which they would have preferred to do, would have involved no last observations carried forward, no imputations. People would be included in this analysis to the extent they had data available to the trial. That is the mixed-effect model and stuff like that.

It means if somebody leaves, he contributes no more data. Now, you could

incorporate into that analysis, data after they stopped taking the drug if you really wanted to. That is not incompatible with intent to treat if you had those data. You only have data on two out of the nine, so that is a little funny to me, but that is the contrast.

Do you take people who truncate their course and attribute actually more points to them than they actually had measured. That is what is wrong with LOCF here. Usually, when you do LOCF, someone drops out at one week, you give them a single value at four weeks, so it doesn't really overweight the thing very much.

Most people are very critical of that way of doing it, but it doesn't sort of overwhelm it. Here, if you drop out at two days, you get 57 days of the same data point, which seems sort of goofy.

But anyway, the contrast is whether you impute, in one way or another, or just take the data you have got and do the analysis, which is what the company wanted to do, but were talked out of it.

DR. LINCOFF: Doesn't the model penalize just on the basis of the patient contributes less data, so if a patient drops out after a few days, there is less data and therefore diminishes the power of the study to show a difference, is that not true?

DR. HIATT: In the mixed model you mean?

DR. LINCOFF: Yes, without the LOCF, if they just used a mixed model, isn't there in effect a penalty for patient dropping out?

DR. HIATT: Dr. DeMets.

DR. DeMETS: The variance structure that you estimate will take into account how much data there is at various points, and in some sense, yes, it accounts for that.

DR. HIATT: Does that help clarify the question?

DR. LINCOFF: Yes, therefore, I do not believe that the LOCF was necessary. I don't think the fact that they are called, you know, anal pain and headache pain mean that they needed to be mixed, so I don't think an LOCF, so does that make

it a yes or no, was LOCF conservative enough.

DR. HIATT: It is a hard way to sort of phrase it though.

DR. TEMPLE: That is a fair summary of what the question is going after. That is what it means.

You could have--I am not even sure of how you would do it, Norm, do you have a way of describing it--you could penalize them more if they dropped out because of headache. I don't know what you would do to penalize them.

DR. STOCKBRIDGE: Like I say, you could have taken the worst rank, the worst observed score, and carried that forward in that person. That would have been worse than their own last value presumably.

DR. TEMPLE: That would be a novel thing to do, but would be done because pain was what you were trying to treat, and they dropped because of pain albeit a different pain? That would be the rationale for that?

DR. STOCKBRIDGE: It has got to be a worse

pain. It is the thing that made you drop out. The thing that brought you into the trial didn't drive you out of it.

DR. TEMPLE: That is because you expect to get benefit for that one. I mean that is what is complicated about it. You don't know people's attitude toward the two pains.

DR. LINCOFF: But with the LOCF, if their last value was actually the improved one, there was a lot of scatter, then, they actually would have had a good value propagated throughout.

DR. TEMPLE: Yes.

DR. LINCOFF: I am choosing to interpret this as should we have used the LOCF to account for this problem and unless someone can give me another interpretation, and for that question, I say no.

DR. HARRINGTON: Using the Lincoff variant here, I would agree with that, that if you were using the LOCF to accomplish the purpose as stated, I would say that it shouldn't have been used, and perhaps that means that there are other methodologies which could account for this problem

of having a different type of pain.

It seems as though there are some other methodologies. Maybe the more conservative one that Norman is going after is not well defined, and that is at least what Mike and I have been struggling with here.

DR. TEERLINK: I sort of agree, but I think that for me, the challenge here is for a symptom endpoint where you have competing risks that are similar to the very symptom that you are investigating.

There needs to be some way to incorporate that into your overall efficacy, and I am okay with mixing bad things and good things when they have to do closely with the same kind of symptom complex, I think. So, in that regard, I think an LOCF or something else that tries to account for that would be an appropriate approach, but whether LOCF is the best approach, I don't know.

DR. HIATT: Dr. Koltun.

DR. KOLTUN: I still don't understand the question that says was LOCF conservative enough, so

I can't answer on this.

DR. HIATT: We have tried to rephrase the question, was the last observation carried forward an appropriate way to address the fact that there are competing kinds of pain. Does that help?

DR. KOLTUN: No. When LOCF is done for the purposes of discontinuance of the therapy, because there is now a competing complication that makes the therapy not worthwhile, then, I think that is probably valid, but what if they discontinued therapy because they felt better, as was suggested in one of the presentations, as was mentioned in one of the presentations?

We are talking here only about those who terminated because of competing headache, correct?

DR. HIATT: For the headache pain predominantly.

DR. KOLTUN: Right.

DR. HIATT: So, your sense of this is an appropriate analysis or not?

DR. KOLTUN: I abstain. I am partially unable to understand the exact subtleties of the

question, while at the same time, I can see the arguments being presented on either side. I don't think I have the competency necessary to make a decision.

DR. HIATT: I think you are challenging us here. When I read the material and came into this, I was pretty convinced that if you are trying to treat a symptom and you cause another symptom, that was not a good thing, and that needed to be addressed.

I thought the sponsor did a reasonable job trying to convince us that the headache pain didn't seem to contaminate the results, and the patients were really coming in for rectal pain, and that is what bothered them the most, and so I don't know if the pains can be treated equally.

So, I am leaning towards agreeing with the sponsor on this one, that we shouldn't overly penalize headache pain and treat that as equal as the rectal pain, because that is what brought them into the study, and that is what they wanted to have treated.

But my bias coming into this was, since it's symptomatic endpoints, we should account for all pain, and a global score would have been nice.

DR. WARNER-STEVENSON: I don't think I can presume to guess whether the headache pain or the anal pain are equivalent, worse, et cetera, so I am going to say no, I don't think that the LOCF should have been used for headache pain.

DR. FINDLAY: Abstain on this question.

DR. KASKEL: I would say no, I think we can always learn more on the natural progression of the disease, and I think we should have had another analysis besides the LOCF.

DR. PORTMAN: Based on modification of the question, I would definitely say no.

DR. PICKERING: Well, I voted for 2.2, so I am not sure if that makes my vote for 2.4 no. I think it probably does.

DR. DeMETS: I think I would say the same thing, but my comment, there is no way to tell if LOCF was conservative enough, because the response is so variable. You just look through those plots,

they are all over the place. So, I don't know, it could have been worse, it could have been better.

The worst rank, I mean that is not a novel idea, people have looked at worst ranks in lots of settings to see whether it matters or not. That would be certainly more severe, but to answer the question, you can't tell whether it is conservative enough.

DR. FLACK: No, LOCF is not appropriate.

DR. HIATT: Okay. Question No. 3. The review team questioned whether concomitant analgesic use could have contributed to differences in the groups. The sponsor has argued that the results are not confounded by analgesic use.

3.1. Do you agree that the results are not confounded? If so, cite the analysis you find compelling.

DR. GOLDSTEIN: Clarification. The only permitted analgesic was acetaminophen, and as I recall, there were eight doses permitted. Was that eight doses in one day or eight doses over the course of the study?

DR. HIATT: Those are good questions.

DR. AZARNOFF: It was eight doses over the first 21 days--

DR. HIATT: Could you maybe answer that question at the mike?

DR. AZARNOFF: Sure. The instructions were not to take more than eight doses over the first 21 days, and no more than three doses per day.

DR. LINCOFF: Was there any assessment of the compliance with that? I mean we routinely collect what other meds patients are on, so that should have been available.

DR. AZARNOFF: Yes, there was a measurement of concomitant medication, and most people complied. There were a few who took other drugs, like diazepam, and so forth, and those were more common in the placebo group than in the active group.

DR. HIATT: Does that help?

DR. GOLDSTEIN: Yes, in that case, my opinion would be that it is not confounded. I just

can't see eight doses in 21 days and no more than three a day as particularly of a product that has little, if any, anti-inflammatory capability.

DR. HIATT: And then the follow-up is for everybody, is there analysis you find compelling that excludes confounding? I might ask as we go around if you could remind us if you included analgesic use in your primary models.

DR. GIBBONS: Analgesic use was used a time varying covariate in the sensitivity analysis. We didn't put in any time varying covariates in the principal analysis, but it was put in as a sensitivity analysis, and it showed no effect.

DR. HIATT: Thank you. That was very helpful.

DR. FLACK: I am much less convinced. I mean I tell people to do stuff all the time, both in trials and clinical practice, and they don't do it, or the doctors have them do something different, so simply saying don't take more than eight doses of acetaminophen really is not overwhelmingly convincing to me of much of

anything.

I think when you are going to look at drug exposure, you really need to look at what they are taking and how much they are taking, and also, you have to remember that this pain is highly variable, and the pain seems to be present one day, gone the next day, and some people are waxing and waning, so they may be taken meds prn around pain episodes.

What I would like to have seen was something where you calculate the therapeutic intensity scores and really get some idea of what people are taking, NSAIDs, as well as acetaminophen, and even though you try to minimize what they are taking, I think you just have to collect it like you don't really know necessarily if they are going to comply or not.

There are some interesting analyses put forward, but the people who I would expect to be on acetaminophen would be the people with the most severe pain anyway, so simply showing that they still have more pain than others, that is what I would expect, because that is the indication for

them taking it.

It is like ACE inhibitors. I can go to a cross-sectional data set and show you ACE inhibitor cause proteinuria. Why? Because proteinuria is why people get them. So, i don't think that this was handled very well particularly given the fact that the effect turned out to not be a very large effect.

Also, I am worried about analgesics, and I am not convinced it's not confounded.

DR. DeMETS: I don't think that I can make a strong case that they are confounded, but I find the evidence compelling that it is not potentially. As Dr. Gibbons pointed out, he did a secondary sort of sensitivity analysis using analgesic use as a time dependent covariate, but as I indicated earlier, those analyses are tricky, so while it is somewhat reassuring, it's not I wouldn't say convincing or compelling, but also I can't argue that there is a case.

DR. PICKERING: I was also not convinced. I agree with John, the fact that the patients

taking analgesics had more pain merely to me indicates that that is the reason why they were taking analgesics, and it wasn't clear before and after they started if there was any effect on the anal pain.

In addition, there was one figure, Figure 5 in the sponsor's presentation that suggests that patients with the most severe headache actually quite a lot more pain than the other group, so I am not convinced that there wasn't an interaction.

DR. PORTMAN: I just don't know that there is a good way to know. I mean I appreciate John's comments and agree with them, but how do you ever know in a study whether patients are taking things surreptitiously. I mean based on the study design, I mean I don't think they would be confounding.

DR. KASKEL: I agree, I don't think they are confounding.

DR. FINDLAY: I appreciate the sponsor's statistical analysis, but in the end, it's pain, and Tylenol is a pain reliever, so I think even though the magnitude of the effect is probably

small, that is just a wild guess based on nothing, but I would say that I would agree with my colleagues that some confounding could have taken place.

DR. WARNER-STEVENSON: I don't think we can rule it out. I am not unduly worried about it.

DR. HIATT: I am in the same position. I think the analysis is helpful, but it doesn't completely rule it out for me.

Dr. Koltun?

DR. KOLTUN: I agree with the last two comments, probably a small effect, but probably not significant.

DR. TEERLINK: I largely agree with the other comments. Clearly, more patients take acetaminophen in the nitroglycerin group, but given the limitations that were placed upon how much acetaminophen they could use, that is partially reassuring although once again, getting at John's point, we aren't sure how that was actually done in practice, but the sponsor did do a good job of trying to keep track of how much acetaminophen was

given for the headaches, and I think as well as you can do it, you are to be commended for trying to do that as best you can.

So, I think it is a potential confounder, but I am not potentially worried about it.

DR. HARRINGTON: I am in the group of uncertainty here, that if you look at the table by the medical reviewer, there is substantially more acetaminophen use in the nitroglycerin group, and given the magnitude of the overall effect size in the trial, where such little perturbations can change things in one direction or another, I don't think we can say that despite the additional analyses, that we are sure confounding didn't take place.

If this was an enormous treatment effect, I would be much confident in saying that I think these additional analyses probably handled the problem. I don't have that level of confidence here.

DR. LINCOFF: I agree, I am somewhat uncertain. I think the additional analyses, as Dr.

Flack pointed out, are not really useful at all, because none of them can sort out cause and effect. But I am reassured by the limited number of doses that patients were supposed to take, and I believe that over only a three-week period, which is where we are really concentrating, that compliance, unless there is some indication otherwise, should have been pretty good.

They did collect the data and short of preventing the patients from leaving their homes to get to drugstores, I don't know how much more they could have done to limit the dosing. So, I think we have to take that at face value. In basis, I am not particularly concerned that it was confounded.

DR. HIATT: Comments?

DR. TEMPLE: Everybody sort of said that they were all agreeing with each other, but I heard two distinct positions. Nobody thought that they were sure they could attribute anything to the aspirin, but some people expressed the degree of concern that a possible confounding had not been ruled out, and others expressed the view that it

just can't account for anything, it might be a little something there, but it doesn't matter very much.

I just want to be sure that we know which the dominant view is here. Nobody thinks it is absolute, nobody thinks it's overwhelming, but I heard some expressions of worry and some expressions of not worry.

DR. HIATT: Maybe we have to go back around the room, why don't we just do that. Why don't you say--are you concerned about confounding, yes, no, and then address the magnitude of effective analgesis.

DR. LINCOFF: So, if I say I am not concerned, does that mean I am assuming no effect? Do have room for middle ground there?

DR. TEMPLE: No. The question is you are worried about the results, such as they are, because of this confounding, that it is enough to make you nervous about the results. It seemed to me different people said different things on that.

DR. GIBBONS: Just a comment that might

help clarify this issue. It is important to note that the treatment effect was larger in those patients who took acetaminophen, so if there was a confound, it would go the other way. So, while we treat it as a time varying covariate, and that is tricky business, as Dr. DeMets says, when we stratified on it and compared placebo to active treatment only in those people who took acetaminophen, the effect in raw units was larger, which flies completely in the face of any potential confound.

DR. HARRINGTON: But isn't that, by itself, a confounded analysis? I mean that's a post-randomization analysis that you are stratifying these patients into groups after they have already been randomized, into whether or not they are acetaminophen takers or not.

DR. GIBBONS: All of these analyses are sensitivity analyses, they are all post-hoc analyses. The confound that was concerned that there was an unblinding, that those people who got a headache, there were more headaches in the active

treatment group, they took an aspirin, so they got better.

In fact, the magnitude of the effect was even bigger when you matched on placebo patients that took aspirin versus including active treatment patients who got better. So, it didn't make the effect go away.

DR. TEERLINK: I thought I understood this, and now I just--so, I apologize.

DR. HIATT: I think the sponsor is telling us that they looked at it. When you put in the model, based on how they captured the data, they didn't see an effect. You are also saying that, in fact, there may have been a stratification around these patients, but suggested it went in the opposite direction.

So, if you actually stick with the evidence that we have here, you would have to say that there is no confounding.

DR. TEERLINK: Were you saying that if you take acetaminophen, you have even less pain with nitroglycerin than if you were in placebo and took

acetaminophen?

DR. GIBBONS: If you compare placebo patient who took acetaminophen, and you also compare treated patients who took acetaminophen, the magnitude of the difference between those two groups is even larger for the groups that didn't take acetaminophen.

In fact, when you look at the placebo patients or the treated patients who took acetaminophen, they had more anal fissure pain. That is why you see a bigger treatment effect. Remember we are always seeing the biggest treatment effect in those subjects who have the worst pain.

The patients who were taking acetaminophen were, in fact, in more pain even after they took it relative to those that weren't. We saw those slides that showed the placebo groups and also the treated groups, and then a comparison between the two.

DR. FLACK: Can you help me understand something? How do you stratify them when their exposure is changing over time, so you may have X

number of visits, but only a fraction will be on acetaminophen, or are you saying that you just stratified the group on the basis of ever taking acetaminophen?

DR. GIBBONS: That's correct.

DR. FLACK: Because that's a different analysis and you don't really need to do that if you are basically handling it the other way, I wouldn't think.

DR. GIBBONS: We tried to beat it to death because we knew this was an important issue, so we did it as a time varying covariate, but then we also did this stratified analysis to show again--

DR. FLACK: You may take one dose or you may take 50 doses and be in that same group is the point I am making.

DR. GIBBONS: That is correct.

DR. HIATT: So, is it fair to say that that is what the data showed, but are we asking whether the designation of taking acetaminophen was a good surrogate for all analgesic use that might have confounded the results, and we don't have

accurate reporting to know that?

DR. TEMPLE: I wasn't getting up to that. I was just getting up to whether you were worried about the acetaminophen and maybe aspirin use, which actually goes slightly the other way.

DR. HIATT: If that is helpful, let's go back and do 3.1. Do you think the results were confounded by acetaminophen and then 3.2 will be the magnitude?

DR. LINCOFF: On balance I believe that the results were not substantively confounded, and as a result, I think the magnitude of effect of analgesics is small enough not to have been important.

DR. HARRINGTON: I have the other opinion. I think that the overall treatment effect is so marginal that I can't be convinced even by these additional analyses that confounding didn't take place.

DR. TEERLINK: I think the treatment effect can have been confounded by the use of analgesics, but it is unlikely to be significant.

DR. HIATT: Dr. Koltun.

DR. KOLTUN: Again, I think there was probably a confounding effect, but for the overall effect on the conclusions regarding the outcomes, it is probably not that significant.

DR. HIATT: Based on those analyses, I am not convinced that there was confounding either, and I guess I am not convinced that acetaminophen would have a big effect on rectal pain.

DR. WARNER-STEVENSON: I don't think we can rule it out, but I don't think it is very likely that it had a big effect.

DR. FINDLAY: Not worried about the effect.

DR. KASKEL: If these patients are in pain and they want some relief, they are going to take it. If you started the design of the study either each group got the same amount of acetaminophen or each group got no acetaminophen, you could have a better analysis, but I think for this study, there is no confounding effect.

DR. PORTMAN: I agree, I don't think there

is a confounding effect, and the magnitude, if there were any, would be small.

DR. PICKERING: I don't think it has been completely ruled out, but if present, was probably small.

DR. DeMETS: I think that is my opinion. I don't think we can rule it out, but given the amount that we see, I don't think it's a major factor.

DR. FLACK: I don't think we can rule it out. These patients' rear ends were hurting, and patients in pain are not necessarily going to do what the study investigator asked them to do, and I just don't think we know, and I am not sure that acetaminophen use is enough of a proxy for all the other stuff that could have been taken to relieve pain including NSAIDs or other potential medications, and if you capture those things, and I think you have to do more than just say they are on them or not on them, we work with data like this all the time, and that's like taking a blunt instrument and trying to shape a piece of china.

DR. HIATT: Comments? Okay.

Now we are up to Question 4. This is a voting question.

Taking all three studies into consideration, do you find the data compelling that there is an effect of nitroglycerin ointment on the pain of anal fissures? Please vote.

Dr. Goldstein, you can comment, but not vote.

DR. GOLDSTEIN: No.

DR. TEMPLE: I just want to make one comment. "Compelling" is a term we often use, but the right term is whether there is substantial evidence. That is the legal standard, so substitute that.

The term, whatever it means in law, is whether there is substantial evidence based on adequate and well-controlled studies, and one of the points they made when they enunciated that standard is didn't have to be perfect, it had to be substantial, but "compelling" is not the term that is customarily used, and it is not the used word

from the law, so I would substitute a different word, a word that is used in the law. But feel free to say how strong you think it is.

DR. HARRINGTON: Earlier, Bill said that the charge of this committee has typically been to look at data from two trials, which have p values less than 0.05, or one that has more robust evidence than that. Is that essentially the background, Bob, on which you want us to evaluate this?

DR. TEMPLE: That is certainly how we have reached the conclusion that there was substantial evidence in most cases, yes.

DR. HIATT: Thank you for those clarifications.

Dr. Goldstein, your light is on. Do you want to say anything?

DR. GOLDSTEIN: I am sorry, it was purely an accident.

DR. HIATT: John?

DR. FLACK: I need some clarification. Are we basically looking at these three, approval,

approvable, not approvable categories?

DR. HIATT: We are not there. That is a vote that comes up later. I guess the question here is the totality of the evidence, is it substantial that there is an effect on pain. We are not yet being asked to take it to the next level because there is some stuff in between.

DR. FLACK: So, we are not really saying what the size of the effect is, we are just saying is there evidence that there is an effect?

DR. HIATT: In the context of a regulatory decision, I think once again, as Bob Harrington just mentioned, the three studies you are looking at, are they cleanly positive, are they marginally positive, or are they unambiguously negative, and do you have enough across the three trials, and you hear people talking about one and one half, or one and three-quarters, or 2.2., the question is, is there enough--

DR. FLACK: This is like saying there is really, really strong evidence or there is not. If you ask me is there really very strong evidence, I

would say no. If you asked is there evidence that there is an effect, I think across the three studies, I would argue yes.

DR. TEMPLE: A little bit. First of all, as the form of the questions indicates, this is not where you should be thinking about whether the effect is big enough to matter. That is for later, it is coming later.

So, it's have they shown an effect on pain. The usual way that we ask that question and get an answer, but I don't want to say there is no alternative, is to look for whether there are two clean studies. They are sometimes called pivotal studies.

That is an unofficial word, it doesn't appear in any regulations, that means, generally speaking, p less than 0.05 by a credible analysis for two studies, but certainly there are cases where we have found one study very persuasive and another study either a little bit, or even one single study of an important endpoint.

What we have generally not done is sort of

looked at three leaners, and said, well, maybe the totality is okay. The usual standard has been adequate and well-controlled studies, generally two of them, that show the effect, but I don't want to build more rigidity into it than there really is.

You can find an example of most things, but usually, what you mean is two studies, significant at 0.05 or very close, or one study that carries you along. That is the usual standard.

DR. FLACK: Here is my answer. Somebody at the FDA told them that or implied that at least one of those first two studies was clean, because they said do a third one, and you might get it approved or whatever.

They picked the wrong endpoint on the first one, but what they found is reasonably consistent across the trials, that there does appear to be a reduction in pain. You can argue about the magnitude, there is certainly differences across the trials.

I would say given all the qualifications, discussions that we had, did they show an effect,

yes, and I will just leave it at that.

DR. HIATT: Okay. David.

DR. DeMETS: Well, I think this is a very hard question to give a simple yes or no answer to. All three studies have issues, which we have been talking about most of the day. For me, they are not overwhelming, and they are not compelling.

Yes, there is evidence that there is an effect of some kind. If you are very strict and formal about it, I don't think they make the criteria that we often use, but I also have to respect and recognize that analysis, which seemed sensible, gives you a lot of encouragement.

So, I guess I would vote either a very qualified yes or a no, and between those two, I suppose I would say a very, very qualified yes.

DR. PICKERING: I basically agree. I don't think you can dismiss all three studies and say there was absolutely nothing there. There clearly is a hint of a signal, but it is very small under a huge amount of noise, and there were also some qualifications like which group you are

looking at, what time point, and how you handle the missing data, all of which makes the overall effect very weak.

I think the other thing is if you look at the aggregate change in both groups, the average pain score went from something like 55 to 35, which is a change of 20 points, whereas, we are looking at the treatment effect, which is perhaps 3 points, so it is again very hard to distinguish it from the noise in the studies.

DR. HIATT: So, that is what kind of vote?

DR. PICKERING: Qualified yes.

DR. HIATT: Qualified yes, okay.

DR. TEMPLE: Can I just make one comment? Our assumption is if you say yes, you think the study, that you have been through all the difficulties and analysis and recognizing what they are, have, nonetheless, come to the conclusion that the studies are credible. I assume that is what it is means. That is probably what you did mean.

DR. PICKERING: Credible, but not pivotal.

DR. TEMPLE: No, no. I mean a yes answer

means that you think that two of those studies meet the test for being an adequate and well-controlled study. I mean the analytic questions are never going to go away, but we have analytic questions all the time, and we sometimes conclude that a study is okay despite that, and that's all right. I just want to be sure I know what you are saying.

My assumption is--tell me if this is wrong--that when you say yes, you mean you are worried about the studies blah-blah-blah, but you think on balance, with all those caveats, they were good enough to be informative. Is that fair?

DR. PICKERING: Well, it's a question about good enough. As I say, I don't think you can say there is no evidence of an effect.

DR. TEMPLE: That is not the question. The question is whether there is substantial evidence of effectiveness. I mean the studies are obviously lean, nobody would say they proved a negative. We wouldn't be here if that were the question.

The question is whether they rise to the

level of being adequate and well-controlled studies that support effectiveness, and that doesn't mean they are perfect, and it doesn't mean there are no questions about them, but it means you have thought about it and resolved the question in a way that says okay, that's good enough, I might have wishes for this, I might have wished for that, but that is a credible analysis, credible study, and it shows that there was an effect on pain.

DR. PICKERING: Minimal.

DR. TEMPLE: Whatever.

DR. HIATT: So, your vote is still yes, and, Dr. DeMets, are you a yes?

DR. DeMETS: I am yes to an outcome. I have no idea what it means, so I guess that is the next question.

DR. PORTMAN: This is all on the positive side, that the drugs works without really talking about its possible safety issues, but I mean I find myself reflecting on whether if this were a condition for a serious--well, serious condition, I would be a little concerned about saying that this

drug really works.

It is more symptoms and the endpoints are very soft, you know, in my mind, but I can't really deny that the three studies together show that there is some positive effect. Certainly, if you have an anal fissure, move to Serbia.

But that bothers me that it was mostly the one country, but I am going to say yes.

DR. KASKEL: Yes, with a qualifier at 21 days.

DR. FINDLAY: Yes, credible studies showing some effect, but not substantial.

DR. HIATT: So, is that a--

DR. FINDLAY: That's a yes.

DR. HIATT: That is still a yes. Okay.

DR. WARNER-STEVENSON: I have some concerns about the magnitude of the difference declining from the first study to the third. I have some concern about the international issue, and I wish we could have seen more analyses of the primary endpoint for different countries, but I will give it a yes.

DR. HIATT: I am going to say solidly no just to kind of really offset I think what I am hearing from my colleagues. This misses my definition of substantial. Of course, there is the suggestion that there is something going on here. I think the miss is not huge, but it just doesn't cross my threshold.

I think that we can discuss what the limitations are at some point in time, but I mean clearly, I think there some underpowering issues, there is some methodologic issues, there is some endpoint definition issues that I think could be addressed, but I don't think this meets my level for substantial evidence.

Dr. Koltun.

DR. KOLTUN: I think I have taken the Question 4 as it is written. I would say--

DR. HIATT: Are you still thinking?

DR. KOLTUN: Pardon me?

DR. HIATT: We didn't hear your answer.

DR. KOLTUN: Oh, I am sorry. I said taking Question 4 as it is written, my answer would

be yes.

DR. HIATT: Thank you.

DR. TEERLINK: So, I would have to say no, and it is because, you know, you are asking for substantial evidence, and I think given that we see relatively minimal statistical differences that are not particularly robust to different sensitivity analyses, that there is the issue of partial unblinding, which could drive these relatively small differences in symptom sizes.

There is a decreasing treatment effect size as we get better trials and more patients involved, and there is this country-specific issue, none of which are, in and of themselves, damning criticisms, but I think all of them end up adding up to taking it away from a substantial.

DR. HARRINGTON: I also will vote no, and really was largely driven by Bob's description or definition of compelling. When I first looked at these data, my thought after looking at all the data was, well, there may be something here, and my comment of there may be something here certainly

doesn't meet the definition of compelling.

DR. TEMPLE: I said the operative term is substantial, not compelling, substantial.

DR. HARRINGTON: I asked for the definition of compelling, you have me substantial evidence, so I linked them together, Bob.

DR. LINCOFF: You said substantial evidence by credible analysis. I am discounting Trial 1, because it wasn't at all one of the endpoints. Trial 2, I think most of us agreed the quadratic analysis was credible and therefore it was significant by at least 0.05.

Trial 3, if we use the prespecified criteria, which I am told they did indeed prespecify headache related to drug, then, the LOCF has a 0.0498.

I think the best analysis, and I was sort of in the minority, was one not using any of the data afterward and no imputation, but even that is significant. So, that leaves us two trials that although borderline, do meet, and since we are not voting on the magnitude effect, just whether there

is substantial evidence of effect, I vote yes.

DR. HIATT: I think, then, we will move on to the next question.

Nitroglycerin ointment administered intra-anally is systemically absorbed - mean bioavailability is 50 percent with wide variability even in a small PK study ranging from 8 to 99 percent. At the extreme, the proposed dose thus delivers 1.7 mg of nitroglycerin in the first hour, substantially higher than the usual anti-anginal dose.

Dosing was erratic in the trials - apparently as much as a 4-fold overdosing based on an FDA site audit. Tachycardia and dizziness were reported in two patients in the small clinical trials, but vital signs were not measured at peak after the first visit. Are there safety issues with the use of nitroglycerin ointment to treat anal fissures?

We did spend a little time on this just before the question. So, I guess we get to start back over here again.

DR. LINCOFF: I think there are potentially safety issues, but I don't think that that is a major issue. There was the issue of there have been very few reports in post-marketing data in countries where this is available. I recognize that we don't appropriately report every hypotensive episode in a patient, but we expect that in patients hospitalized were receiving nitroglycerin for angina, and I think if patient syncopized or had other major hypotensive-related complications when they are getting something for anal fissures, there might be at least some of those reported.

So, I think that the lack of reports in post-marketing is somewhat reassuring, and although this is a theoretical argument, I think that the very real, what sounds like terrible discomfort associated with the anal fissure, that this is a relatively mild theoretical safety issue, and I don't think it's an overwhelming problem.

DR. HARRINGTON: I think that there may be safety issues, the dizziness reported in the

combined safety analysis was I think almost 4.5 percent. We didn't have good clarification on what might have been associated with that including pre-syncope, syncope, but I am reassured by the fact as we talked particularly with our colorectal clinical colleagues, that this is mainly a disease of the young, and it is likely that these safety issues can be overcome through perhaps some education, et cetera.

So, while I think there may be safety issues, I agree with Mike that they are probably of a modest level.

DR. TEERLINK: And I agree with that. Yes, I do believe there are safety issues, but I do not believe they are severe. I think there is clear evidence of increased headache and probably all the sequelae that go along with the hypotension, but we don't know that, and we will see that perhaps if this does get approved in greater numbers of patients.

So, yes, there is a safety concern, but no, probably not severe.

DR. HIATT: Dr. Koltun.

DR. KOLTUN: I basically echo that. One of the issues that I forgot to ask when the data was first presented was, was there an exclusionary criteria for individuals who had known heart difficulties, cardiovascular problems, because I know in my personal practice those patients almost uniformly can't tolerate this drug.

As for it being a youthful disease, it is not always a youthful disease, but generally speaking, I agree with what has been stated so far, that this is a relatively safe drug when administered as instructed.

DR. AZARNOFF: I can tell you there was no criteria that excluded subjects with cardiovascular disease.

DR. HIATT: Okay. There was no criteria to exclude cardiovascular disease, but you did exclude phosphodiesterase inhibitors.

I will just echo that, too. I think the things that were compelling for me: short course of therapy, relatively young population, lot of

known data on safety with this compound, but if there is a risk, I think the absolute risk is probably very low, so for someone who spends some time thinking about safety in these studies, I am not terribly concerned.

DR. WARNER-STEVENSON: I think there is clearly a concern that would need to be expressed very explicitly to each patient perhaps more so than is currently being done, however, I think the rate of severe events would be so low, you would be very unlikely to capture it in these trials, and one would have to look for it later.

DR. FINDLAY: I agree with the comments made so far, particularly with respect to the short course of therapy, I think the safety profile is probably in the realm of acceptable.

DR. KASKEL: I agree, small safety issues.

DR. PORTMAN: I am concerned. I don't think we know enough particularly about what happens to the blood pressure, and I think that people need to be able to make an informed consent, and I think that while they may not need to go back

and do another study, there should be some acknowledgment that we don't know and in particularly the elderly, people need to be concerned about it.

DR. PICKERING: I agree that it hasn't been really adequately evaluated particularly as has been mentioned with the phosphodiesterase inhibitors, and I think in most of these subjects, it would not be a major concern, but I think it has been under-investigated.

DR. DeMETS: The short-term usage and limited number of patients, I think there is probably--it could be a safety issue--but it is probably small, and these studies are not going to identify it. It is going to take more observational data.

DR. FLACK: I think the safety issue is probably not great. Also, though, I am disappointed in the level of data on a very easy to obtain endpoint like blood pressure, which could have been presented in a much more informative manner.

I think the problem you potentially run into this is when you start using this drug, and people on drugs like alpha blockers, or people who have been overdiuresed, or maybe even people with diastolic heart failure, and trust me, I have those patients in my practice in their 30s and 40s, so you don't have to be old to get those problems.

But in an absolute sense, I don't think it's great, and hopefully, the sponsor has got the data. I could never get a clear answer about how they recorded the data, but telling me the diastolic pressure dropped 20 mm of mercury as opposed to something a bit more conservative, and probably closer to 10, which is probably more standard for orthostatic hypotension, and that is another group that could really get hit with this drug.

I think it is just an incomplete picture, and maybe they just didn't think about it, but I still have concerns, but not great.

DR. WARNER-STEVENSON: I just wanted to add, listening to you, as well, I do think it bears

considerable emphasis that we do not know how this is going to interact with people on any of the blood pressure medications, and I think one would want to think very strongly about excluding those patients until we knew more.

DR. HIATT: Dr. Goldstein, any comments?

DR. GOLDSTEIN: The only comment I would make would be to point out the liberal use of the term speculative, potential, et cetera. Without question, a drug like nitroglycerin would have side effects for somebody somewhere in these uses, but balanced with the evidence for effectiveness, et cetera, I don't think they would be very large. They would most of the time be able to be handled by further both post-marketing and other labeling and other well-designed techniques.

I just don't think they would be a problem, particularly with the background of a century or more of the use of this agent. It would be simply a matter of collecting the information as it comes in. What we do have is the foreign experience, and the English, in particular, are

pretty good at this from a variety of ways, and we have heard none from there.

DR. HIATT: The next question gets kind of interesting.

Independent of the need to show net benefit exceeding risk, which of the following factors, if any, influence whether or not the size of a treatment effect matters for regulatory decision-making:

Benefit is a reduction in major clinical outcomes.

Benefit is an improvement in functional status. I am assuming that is like an SF36 functional status instrument.

DR. TEMPLE: I think angina maybe or heart failure and Living with Heart Failure Scale, something like that.

DR. HIATT: Exactly, for functional status.

Benefit is an improvement in global patient assessment. Since we are on a roll here, do you want to help characterize what that is?

DR. TEMPLE: Well, there are patient globals and patient-reported outcomes that try to look at the patient's whole state, including the physical state, mental state, the sociopolitical state, and a wide variety of other things as distinct from a single isolated symptom like pain.

I am not sure how much you have to get into the larger question that we are going to get at someday, but we are really obviously mostly interested in how you feel about an effect on an isolated symptom, which is what we are talking about here, so feel free to consider the other things, but in the end, that is what we are really dealing with.

DR. HIATT: All right. Where do we begin?

Dr. Goldstein, do you want to comment on what matters for regulatory decision-making?

DR. GOLDSTEIN: I think in various ways, one could make a case that all of these can matter under certain circumstances, but I would think an improvement in functional status would be my number one choice.

DR. HIATT: I think just to frame this, which of the factors would influence on whether or not the size of a treatment effect matters? So, if you are doing an event-driven trial, and one more person lives on drug than placebo, does that affect size matter out of 10,000 relative to three points on a pain scale? I think that is how we are looking at this.

DR. TEMPLE: That's right. As I indicated in my comments, we, for the most part, don't actually look at how big the effect is as long as you can demonstrate it. This sort of gives a range. A decrease in mortality that was small might be considered valuable no matter how small it is if you could detect it.

Maybe the same thing on heart failure. Some people can go about their business better, maybe it doesn't matter how many can, you wouldn't care. So, it is an attempt to look at these various outcomes and where the actual size of the effect starts to matter.

As I said, we historically have not

particularly worried about that, and the question this raises is are there some cases, like this one, where you are just talking about a particular kind of pain. Some people have said it's a very important kind of pain, and does effect size start to matter for some of these. That is what this is the introduction to.

DR. HIATT: Dr. Goldstein, comments?

DR. GOLDSTEIN: The effect size probably would have an effect in terms of--and I am unclear as to what is meant by "reduction in major clinical outcomes."

DR. HIATT: Imagine an event-driven trial where your outcome is myocardial infarction, stroke, and death, or say it's a mortality study where death if the primary endpoint.

DR. GOLDSTEIN: Well, it certainly would matter there, yes, it certainly would.

DR. TEMPLE: No, I think what your answer is, is that it doesn't.

DR. GOLDSTEIN: Right, exactly.

DR. TEMPLE: What you are saying is any

effect no matter how small on an endpoint of that magnitude counts.

DR. GOLDSTEIN: I meant the other way around.

DR. HIATT: Not to dwell on this, but then do you think the size of the benefit matters for an isolated symptom?

DR. GOLDSTEIN: No, I do not. The size, you know, that, it wouldn't matter. A global patient assessment, well, again, global, I am troubled by the word "global" in a situation like this.

I think in the case of functional status, these people can't function effectively if they are in pain. If they are constantly in pain because of an anal fissure, it is difficult to function, so I think there, too, it wouldn't matter.

DR. HIATT: There, too, the magnitude of the effect wouldn't matter, is that what you are saying?

DR. GOLDSTEIN: If there were little effect, and they were still incapacitated, well,

you know, it wouldn't really do you any good. I think basically, functional status, I think is crucial here.

DR. HIATT: In this particular indication.

DR. GOLDSTEIN: In this particular indication, that is what I am trying to say.

DR. FLACK: I would argue that the effect size in all of these matters, and I would clearly say for an isolated symptom, it matters, but I would even argue for major clinical outcomes it matters, because you can have studies that show one, two percentage point differences in an absolute sense.

In all practicality, there is essentially no difference, but you study 15-, 20-, 30,000 people, and you come up with statistical significance, and that has been seen with thrombolytics, it has been seen with blood pressure trials, and you end up with people running around making claims about this is better than that, and the next time you do the study, it may be just the opposite or there may be no difference, so I think

that at some point, we are going to have to get smarter about specifying the magnitude of benefit.

I also believe even for symptoms, that it is important to have some magnitude of benefit, particularly if you are going to take risk treatment. I mean we treat headaches, we treat knee pain, we treat in osteoarthritis, degenerative joint disease, RA, so I don't have a problem treating the symptom.

We do a lot of that in medicine, but I think that when we start prescribing therapy to people, and all of them have risk, that there ought to be more than just a nominal increase in benefit.

So, I think effect size matters in certainly for isolated symptoms.

DR. DeMETS: Well, first of all, estimating size and effect, in whom are you trying to estimate this? The populations we study in trials are usually not representative of anything.

They are just what we have in our hands, so we can fuss a lot about the precision of that estimate, but it doesn't apply to the general

public or even patients that fit the entry criteria, because that is not what we study. We study something that we have got.

So, with that caveat, it is hard to figure out what it is we are estimating. I think that the net benefit and risk matters for all of those categories, increases are you go down.

In this case, in this particular study, I suppose, you know, it would be mainly on the issue of isolated symptoms, but, in general, we can reduce stroke, mortality, and death, very expensive procedures or devices, and you have to ask--which have their own morbidities--we have to make some judgment about what is a big enough benefit relative to the risk.

So, I think they all are important, but certainly for things which are kind of the symptom, I think the benefit really needs to be substantial relative to the risk.

DR. TEMPLE: Can I ask a question? David added something at the end that is different from what we are really asking. You said it has to be

substantial relative to the risk.

The question we are trying to pose here is assume there is no big deal risk, okay, and it isn't--maybe this was poorly phrased--it isn't whether effect size matters, of course, it matters. The question is does the effect size have to be of a certain magnitude in order to be considered evidence of effectiveness. That is really what the question we are asking is.

DR. PICKERING: Well, certainly they matter. I think it is very hard to quantify. One way of looking at it is the number needed to treat to get a given effect size, which I guess is saying it in a different way, but some of these objectives, the first one, major clinical outcomes is pretty much a hard endpoint, and they sort of get softer and more subjective as you go down the list.

With regard to the isolated symptom, I mean certainly in this context, pain is obviously very important. Perhaps one thing you could do would be to get some estimate of what people who

have the symptom regard as a meaningful difference. I don't think we have heard that.

I mean we are talking about a 3-millimeter difference in this study, and I don't think any of us have any feel at all for what that really means in terms of daily life, but it is something that you could potentially explore.

DR. PORTMAN: I agree with Tom, but assuming that the study is a well-done controlled trial, and it has a statistically significant result. Then, the question is if it has to be of a certain magnitude, then, based on the talk that you gave previously, aren't you going to inhibit companies from really developing new drugs.

I mean that is what you have been asking for is a statistically significant controlled trial, and it shows that effect. Therefore, the company has gone through the expense of doing all of this, and you want to encourage them to do that.

Maybe one particular drug may not be as effective as another, but I still think that that would be where it would warrant an approval, and so

on.

DR. KASKEL: I think what has been said thus far, I agree with. It's of descending order of importance. Obviously, the major one is the reduction in clinical outcomes.

I would like to add, though, that I think what is lacking here in the study is some data on the natural history of progression of this condition, and a questionnaire for long-term follow-up whether the patient drops out for symptoms or not is recurrence, and what happens to them 18 months down the line, we don't know, and that would be important.

DR. FINDLAY: Yes, the magnitude of the effect matters and should be taken into consideration more often by the FDA.

DR. WARNER-STEVENSON: Yes, I think the magnitude of effect matters. For the major outcomes, however, the lower limit is very low, partly because, in general, when there is an effect on major outcomes, that is the tip of the iceberg

and you are probably affecting a lot that you are not measuring.

As soon as you move down to the others, I think the effect becomes very important, but we have to consider two things. One is the number of patient affected, and then two is the amount of effect per patient affected.

I think it is very important that we begin collecting data in a way that allows us to distinguish those two. Functional status and isolated symptoms, I think are very compelling.

The global patient assessment actually worries me a little bit, because I want to know what we are doing makes the patient better. I think it is a little harder to understand since narcotics probably would work really well for global patient assessment, but I am not suggesting that we bring that before the committee.

DR. HIATT: I have got to take a very different approach on this one. I say no, it doesn't matter. Has it mattered it in the past, it shouldn't matter today, and I think the reason I am

saying that is from a regulatory perspective, that I think the marketplace and people that make policy, and people that pay for therapies should decide if the effect matters or not.

I think that the job here is to say does the evidence support that the effect be it placebo or whatever the question is, so I mean I believe that those are relevant, important questions, but I am not sure the process that we go through should necessarily be driven by that.

I think the challenge here, frankly, should be stated slightly differently. It is not whether an isolated symptom was better on drugs than placebo. It is why should we care.

I think the problem we are having here is interpreting an isolated symptom relative to placebo in the context of did it change anything else that might have been relevant to that patient in the healthcare system and everything - less surgery, better global functioning, net reduction in total pain burden.

I think the challenge here is that we

don't know how to interpret a positive study if these studies indeed are positive. I think that in symptomatic trials, for example, in other cardiovascular indications, we will have an exercise endpoint, which is an objective measure, we will have quality of life scores or functional status scores, but I have always interpreted clinical relevance based on whether the patient can perform better on one measure and then there is concomitant measure that supports that performance if they recognize it's improving their symptom.

Just to focus on a single isolated symptom, in my mind, doesn't tell us that much. So, I am struggling not with whether this should be by a lot or a little, because the default would be, well, it if it's one pain syndrome, and this other pain is going on from the therapy, it should be really good to be approvable, I don't think I would go there.

I believe that it should still be placebo by a statistically significant, unambiguous result, and the interpretation is what is challenging. So,

I vote no.

DR. TEMPLE: Can I just comment? There is some movement toward a much greater likelihood of looking at patient outcomes other than the very thing itself. For example heart failure trials all include a Living with Hearth Failure Scale in addition to exercise when they are done.

We see a lot of them in Oncology. They are not easy to win on because the particular thing you are treating has to be fairly dominant in the person's life. Now, in this case, maybe it is, but in other cases, it may not be, so that you can have a treatment that makes your migraine headache shorter, but if you are not having a lot of migraines, it may not show up as that.

People have tried fairly hard to be able to get those claims because payers want them. I think that's the reason anyway, and it's not entirely easy.

I guess the other thing I want to add again is if we really are interested, as this side of the room was, in effect size and making that a

criterion, we really have to build it into our null hypotheses. Otherwise, you are just looking at the point estimate.

That is not reasonable, that is not a measure of effect size in the usual way, or let me throw it back to you. What is the right measure of effect size, the mean, the range, the thing that is just at the lower bound of the 95 percent confidence interval?

We have got to decide all those things if we are going to start to use them.

DR. HIATT: Yes, there are consequences to that decision or that recommendation. I totally agree with that.

Dr. Koltun?

DR. KOLTUN: These are four questions. I think a lot of people have been talking about, you know, heart disease and other examples, and such, but we are talking about nitroglycerin therapy and these studies specifically, and, you know, I am trying to be objective here, but I am a clinician, and I am a clinician that takes care of this very

problem. I am a clinician that uses this drug, and it's getting to be kind of late, so I am going to try to cut to the chase here a little bit.

If you look to these four things that are stipulated in Question 6, which one applies in this scenario, well, basically, No. 4 applies, benefit is an isolated symptom.

What has been shown here or suggested, I think suggested strongly, and something that I believe, is that there is a rapid improvement in pain when nitroglycerin ointment is used for patients who have a chronic anal fissure.

However, I do not think, and I do not think these studies, in any way show that any of the other things that are listed in Question 6 have been proven.

There is no alteration in major clinical outcome even though people are talking about 35 percent incontinence for surgery, there are many studies that have shown that using nitroglycerin ointment like this in chronic anal fissures, in fact, has a high failure rate for healing of the

fissure, and remember that that is not what we were studying here. Remember that all we are talking about is anal pain.

We never talked about the fact that these studies didn't show healing of the fissure. In fact, in the clinical scenario, what I typically find for these chronic anal fissures is that, in fact, the patient feels much better very quickly, but that they have a high relapse rate, and that high relapse rate happens two and three and four months later, and then you get them back on nitroglycerin ointment and they feel better, but they don't heal their fissure and they have a relapse rate, and they do have functional compromise because of the pain, but are you going to keep them on nitroglycerin for the rest of their lives?

No, you basically resort to surgery, and surgery actually is quite safe. So, my point in going on like this is that this drug is used, and it is probably a very valid therapeutic that should be in the hands of clinicians.

However, the studies that have been shown here today mirror exactly what we find in the clinic already, because this drug is being used a great deal off label, and that is, those fissures that would have healed anyway, heal when you give them nitroglycerin, and the difference is that those patients get a rapid improvement in their pain as opposed to the more indolent improvement that you would get with other conservative management, but that those fissures that are really chronic and really disabling and really bad actually do recur because they have fundamental physiologic reasons for being there, that are more accurately and effectively addressed by the surgery, which is not ordinarily disabling.

But the drug has value, and the drug's value is No. 4, as correcting an isolated symptom, i.e., pain, in a relatively acute fashion. However, it does not cure the fissure, it has a high relapse rate in the chronic fissure situation, and it does compromise functional status, and therefore, it is effective, but for long-term

functional status and long-term global assessment, my personal feeling, and there are studies in the literature that support me, it is not really that effective for really bad fissures. Okay?

DR. HIATT: Thank you.

Dr. Teerlink.

DR. TEERLINK: In some ways I actually will focus just on the regulatory decision-making aspect of it, which to me actually makes this question a bit nonsensical, because I don't think we ever make decisions based purely upon benefit.

You can have a drug that shows improvement in one type of symptom, but if it had a horrible safety profile, I think we would be reticent to regulatorily approve it, so given that context, I think that there is a gradation of need to have treatment effect as you go down to less strong endpoints, but whether that is a regulatorily defined thing or whether that kind of just plays into our psychology as we make the decision, I can't say. I don't think it is a regulatorily defined aspect. Did that answer?

DR. TEMPLE: Not quite. Does that mean if you show that it is good enough?

DR. TEERLINK: Show it--

DR. TEMPLE: The previous questions, No. 4, are have you shown an effect on pain. The usual response by us would be oh, okay, you work, and we wouldn't say is that effect big enough to be counted even.

That is what we are asking you about here, is this a case, or are there cases, where the effect should be at least 3, or 2, or something, should we try to define a minimum important effect or minimum important difference.

Leaving aside all the problems that arise when you are looking at mean effects versus effects in individuals, big difficulty in doing it, but that is the question that is raised by this, because of at least a perception by some people that the effect size is very, very modest. That is what we are asking about, how much should it matter.

The thrust of all that was if you improve

survival, there is nothing too modest. I mean you can decide not to pay for it if you are payer, that is somebody else's business, we don't care about that, but any effect on survival counts if you can show it, and if it takes a 40,000-patient study, fine, there have been 40,000-patient studies to do just that.

But then as you get down to an isolated symptom, you might ask a different question. You might say is this good enough to matter, not because the drug is toxic and it is going to hurt somebody, that is a different question. We would always worry about that.

But even if you are not worried about that, is there an effect size so small, whatever that means, that we really shouldn't approve the drug.

DR. TEERLINK: It you are asking me whether size matters, I would say that yes, in certain circumstances, and as you get to softer endpoints, it is more--yes, I think actually, there is.

I would say I think the importance here is whether it is clinically meaningful, so yes, you can show a difference on these VAS scales. In multiple series, at least for dyspnea and other kind of pain indices in other trials, you need at least 5 millimeters to be something that kind of distinguishes one patient from another patient in terms of whether it really is a difference.

So, yeah, I think there is a level that is clinically meaningful for symptom things.

DR. TEMPLE: And that if the effect size on average, I mean a further question is how to implement that, because when you look at averages, that is not what happens to everybody, and we don't have really great methods for looking at what the effect is in individuals.

I mean if you did multiple crossovers, maybe you could get that sort of data, but the usual randomized parallel trial doesn't give you that. You can show the distribution of responses, which is a clue. It is not exactly the same as what the drug did for people, but it is a clue.

Anyway, that is what this question is about.

DR. HARRINGTON: To answer the specific question, maybe not your second question that you just posed, I do think that the size of the effect matters, and I think it matters more as you go down this list here perhaps where the outcome is a bit more ambiguous or less well defined.

Where the hard work is going to come is to define what those minimally important differences are, how to measure them, how to analyze them, et cetera, and I will stay away from answering that part of the question, but I do think it matters.

DR. LINCOFF: I actually do not think it matters, I think for any addressable, be it either clinical outcome or pain that is worth treating, so obviously, there are some things that we have talked about that may never be worth treating, but assuming it has some significance, and pain certainly has significance even though it's the bottom, if you don't have a risk that outweighs the benefit, I think it is arbitrary to define a

benefit that is useful.

If it's perceptible, and the patient can tell the difference enough to show up as a significant result in a trial, then, it is having some impact on some patient.

There may be value decisions involved, can we afford it, and again there are other more risks, but assuming those issues aren't really--the risk is aside, that we don't have that problem, the other issues are not really regulatory.

So, I think in the pure sense, should something be approved on the basis of regulatory approval, I think if it has an effect and an endpoint that someone cares about, they may not care about as much as death, but still someone cares about it, that the effect size is not the important issue.

DR. TEERLINK: But for many of these endpoints, when they have looked at them, they have tried to ask, the clinicalness is where you have tried to say what is the difference that a patient can perceive, so you are actually saying, you are

actually in some ways agreeing that there is a minimal amount, and that has to be a perceptible difference, and in many studies, a 1-millimeter or a 20-millimeter--I won't take on any specific thing--may not be perceptible to the patient, but it may be statistically different.

DR. LINCOFF: But presumably, you wouldn't see that difference in the endpoint if it couldn't be perceived, if a patient couldn't say I got 5 millimeters better, it didn't feel 5 millimeters better, he wouldn't change. I realize this isn't one crossover design, but presumably, it is perceptible if it's different.

DR. TEMPLE: No, you can see changes in a group of people that each individual might not perceive as a change by himself. I don't think it's obvious that that is not of value. If you show the distribution of results, and more people get an improvement of 10, 9, 8, 7, 6 than in the untreated group, you might interpret that as being a benefit. More people are getting to the place they want to get even though any individual would

be very hard put to know for sure whether he just got better because time passed or got better because of the drug.

DR. HIATT: There could be population benefits to that, too.

DR. TEMPLE: Yes.

DR. HIATT: It may be imperceptible in individual case level, but there may be societal good.

DR. TEMPLE: Just as a typical example, this isn't a secret. In depression, the difference on the HAM-D score at the end of the study, the Last Observation Carried Forward, of course, between the treated and the untreated groups is 3 points on the HAM-D.

You know, there is other kinds of data showing maintenance and stuff like that. That is for all the drugs we know and think are effective when they win. About half the studies can't distinguish drug from placebo at all. Yet, I think very few psychiatrists would say that antidepressants don't work.

So, maybe it is something about the trials we do, I don't know, but that puny three-point difference is not very obvious, and yet that is how the trials come out. It is clear that that is different from the placebo group, that is why they win.

So, this remains something of a mystery. That's why we are posing hard questions. It is not quite clear what the individual results in the effect size means in these settings, whether something about trials tends to shrink them, I don't know.

DR. HIATT: Okay. I hope that was a helpful discussion.

DR. TEMPLE: It was actually.

DR. HIATT: Good. Dr. Koltun, you are still on, right?

DR. KOLTUN: Yes.

DR. HIATT: I was just informed that if you were to hang up, we are not going to recover you, the line goes off at 5 o'clock, and it is about that now, so be careful not to hang up your

phone. That's my message.

Let's try to move on. We have a number of questions actually that we need to get through here.

Question 7. Does treatment of anal fissure pain belong to a class of indication for which the effect size matters? If not, go to No. 10. I am hearing most people aren't going to go to No. 10?

Let's go around the room and just quickly respond to this one and then keep going.

DR. LINCOFF: I said it doesn't matter.

DR. HARRINGTON: I think it probably does matter.

DR. TEERLINK: Yes.

DR. HIATT: Dr. Koltun?

DR. KOLTUN: Yes, it does matter.

DR. HIATT: To be consistent, my answer is I don't think it matters either.

DR. WARNER-STEVENSON: I think it does matter, but I don't know how many individual patients experience a lot of improvement, so I

don't know what the effect size is from this data.

DR. FINDLAY: Yes.

DR. KASKEL: No.

DR. PICKERING: Yes.

DR. DeMETS: Yes.

DR. FLACK: Yes.

DR. HIATT: I think that means we are going to move on to No. 8. These are actually very interesting, so let's try to give these a go.

Question 8. The instrument used to assess effectiveness in these trials was a 100-mm Visual Analogue Scale. In Study 3, mean response in the placebo group is shown in the figure below, no imputation. Page 19 of the sponsor's briefing document.

8.1. Since subjects had to have some minimum pain score to get into the study, some of this effect is regression to the mean. Can you estimate how much is regression to the mean and how much is the natural history of the disease?

John.

DR. FLACK: Jim Eaton years ago showed me

how to do some of this, but I don't remember. I am not sure it matters as long as it is less favorable in relative terms compared to the treatment group, but I honestly don't know how you would just tell right offhand regression to the mean and natural history of the disease per se.

DR. DeMETS: Yes, I think the key thing is to have the comparison, because that's the way you can find out if they are both moving at the same time. If you had to make a guess about regression to the mean, you would assume it would happen relatively quickly, so somewhere in the first few days is probably regression to the mean and the rest perhaps treatment, but I wouldn't want to get too precise about that. The bottom line is you get two groups to compare.

DR. PICKERING: I would say between 10 and 20 on the grounds that from day zero to day 1, there was a universal or an average decrease of 10 in the score, and the overall score was about 20 or so.

DR. PORTMAN: I don't know how you could

tell.

DR. FLACK: Abstain here.

DR. WARNER-STEVENSON: I don't know how you could tell, but I would just emphasize sort of the issue of eligibility creep. When someone has to have a score that is this high to get into the trial, the scores are going to be higher for the first day. I think that will go away relatively quickly.

DR. HIATT: Yes, I agree. I don't think you can tell either. It is probably a combination of both. What I heard about natural history would suggest that there is improvement in this symptom over time, so part of this is probably related to natural history.

Dr. Koltun.

DR. KOLTUN: I wouldn't know how to measure that short of, you know, devising a way to look at it.

DR. TEERLINK: I think the first part of the curve is pretty clearly regression to the mean given how dramatically everybody gets worse on the

first day. Suddenly, everybody, you know, a lot of people have less than 35, so that certainly is part of it. In terms of the rest of the trial, I am not sure.

DR. HARRINGTON: Yes, I am also not certain. My guess would be that the early improvement is regression towards the mean, but after that I am not sure.

DR. LINCOFF: I agree with the previous comments.

DR. HIATT: Okay. Then, turn the page. The figure below shows the mean effect in the placebo and active treatment groups in Study 3, with no imputation.

8.2. How large is the nominal treatment effect, active minus placebo? How does it compare with the effect seen in the placebo group?

Let's start back over here.

DR. LINCOFF: Well, obviously, effect size varies over time, so it looks like it reaches a peak around days 13 to 16, and, you know, you can estimate numerically, it looks about 5 to 7.

Again, I think the regression in the mean is shown at least as a minimum on the next page, where all those ones that have sudden dropped, I think those are the minimum of the regression of the mean.

So, if you have to discount some of this difference and perhaps start from a day or so late, you know, the magnitude of the treatment effect is maybe a third of the magnitude of the reduction seen spontaneously, so it is not trivial at its peak.

DR. HARRINGTON: I agree with Mike's interpretation.

DR. TEERLINK: I agree although I discount the first day, so that I could see down to about 40, so 40 would be the baseline from which I would work.

DR. HIATT: Dr. Koltun.

DR. KOLTUN: I would agree, you know, it's a relatively minor difference.

DR. HIATT: These data confirm my view of the world that it is a good thing to go to placebo-controlled trials and if you are on the

placebo group. So, I think the point here is obvious, that the treatment effect is relatively negligible to whatever the regression to the mean natural history course of the disease is.

DR. WARNER-STEVENSON: I think there is clearly an effect and what I don't know is if some people got a lot better and other people didn't get better at all, but usually, when there is some difference there, some people had to improve a lot.

DR. KASKEL: I agree with what has been said.

DR. PORTMAN: Me, too.

DR. PICKERING: I agree.

DR. DeMETS: I don't think I can add anything. As was said, the key things are in the two groups to compare, and there it is.

DR. FLACK: I agree with all my smart colleagues.

DR. HIATT: Well, in that case, we will ask you the next question.

The figure below on the next page shifts the placebo and active group curves slightly and

adds all of the observed data. Along the bottom now runs the discontinuation rate in the two groups, and we have seen this.

A patient, regardless of Cellegesic, is generally going to feel better over time. Is a patient apt to perceive the contribution Cellegesic makes?

DR. FLACK: That is a several hundred million dollar question. I think some of them are, looking not just at this, but some of the times to percent improvements and all, I think they are likely to perceive it as a shorter time to getting to a certain level of recovery, and I think it is sort of buried in here, so my answer would be yes. Is it 100 percent? Probably, not, and we don't know which subgroups benefit most, but I would answer yes.

DR. DeMETS: Well, I wrote down my list of things. I had a whole page of things I was concerned about. This figure kind of represents the one at the top of my list. I just don't understand, or very comfortable with this outcome

variable, what the kind of noise and variability that this represents.

So, I suspect that some patient can perceive it, but most, probably not.

DR. PICKERING: Well, I think the patients are going to attribute the change in symptoms to the Cellegesic. Whether or not they are reliable or not is highly questionable, but I think just the fact of introducing some treatment that is thought to be effective has obviously had a big effect on their symptoms, and I think in real life practice, it is very hard when you are talking about relief of symptoms, to always distinguish the placebo effect, which is seen in practically every treatment we give from the physiological effect of a drug.

DR. PORTMAN: I mean clearly as a group, since this is a symptom score, more people perceived it on the nitroglycerin than not, but what an individual patient would perceive, I don't think you can know from this.

DR. KASKEL: I agree you can't

underestimate the placebo effect.

DR. FINDLAY: I agree. I think it's clear some patients will perceive the difference and get some benefit, but it may be close to 50 percent or it may be less.

DR. WARNER-STEVENSON: I agree. Very few patients perceive an average benefit over a population, and also, once we prescribe the drug, if they get the placebo benefit plus the active drug benefit, they can't tell the difference in those either.

DR. HIATT: I will just echo that. My earlier comments where I was not convinced the data were substantial for approval, so therefore, I wouldn't want to interpret it any further than that.

I think it's impossible from this figure to predict individual patient responses, but it is fairly convincing the drug causes headache.

Dr. Koltun.

DR. KOLTUN: It's a small effect, but a real one.

DR. TEERLINK: No.

DR. HARRINGTON: There is obviously a small overall effect, which therefore, for me, makes it very challenging to sort out if there were individual patients who would be able to perceive the benefit using these data.

DR. LINCOFF: I think individual patients perceiving benefit will be difficult, but clearly, some patients, there are a proportion of patients who seem to be much better, but, of course, how can they perceive that as compared to their cohorts who got placebo, so I am not sure how you can tell if a patient would perceive the difference, but on average, more patients will be more better on therapy than they would not be. That's how I interpret this figure than without it.

DR. HIATT: Any other comments on this question?

[No response.]

DR. HIATT: Question 9. The sponsor presents an analysis, backgrounder pages 39 to 42, to show that the effect of Cellegesic is larger in

upper quintiles of baseline pain score. Compare this with an analysis performed using a 10-mm-wide moving bin, show in the figure below.

Does everyone know what a 10-mm-wide moving bin is?

DR. STOCKBRIDGE: I think I can explain. What was done here was, first of all, the vertical lines, the dashed vertical lines, mark the boundaries between the various quintiles, and the bottom, the x axis represents the baseline score.

So, what was done at every point from 35 to 95 was to take that nominal baseline value plus or minus 5, and assess what the change was from baseline to the end of the study in pain score, so you could look at it is really a smoothed version of the treatment effect by baseline score.

DR. HIATT: So, just to interpret what you are trying to show us here, at the baseline scores going from about 35 to 55, the groups overlap from wherever this starts, 55 to about--

DR. STOCKBRIDGE: That is the first three quintiles, that boundary there is the top of the

third quintile.

DR. HIATT: And then the next quintile, you have separation, and then the one above that, the curves seem to come together.

DR. STOCKBRIDGE: Exactly.

DR. TEERLINK: And just to clarify, this is for the 21-day?

DR. STOCKBRIDGE: This is the effect at 21 days.

DR. FLACK: Does this include all follow-up measurements, or is it just the final study measurement?

DR. STOCKBRIDGE: This is all--I can't remember what I did.

DR. TEMPLE: But basically, Norm, it's the fourth quintile where there is that separation?

DR. STOCKBRIDGE: Right.

DR. TEMPLE: And the fifth quintile where the separation comes back together, would that be correct?

DR. STOCKBRIDGE: That's true.

DR. TEMPLE: So, in the fourth quintile,

there is what seems to be a large difference, and that's 20 percent of the population, right, since it's a quintile?

DR. STOCKBRIDGE: Right, exactly.

DR. TEMPLE: There appears to be, you know, the placebo group is getting an effect of somewhere in the neighborhood of minus 20, and the treatment group is getting an effect somewhere in the neighborhood of minus 45 to 50, but then in the even sicker group, the people who start out in the 70s, 80s, and 90s, you really don't see that kind of difference anymore.

DR. STOCKBRIDGE: Right.

DR. TEMPLE: So, there is that middle group who are sick, but not too sick, I guess.

DR. GIBBONS: Just a brief comment to help explain this plot. This is kind of a whacky plot. The problem with these kinds of plots is that, number one, they make up a lot of data, so what you are looking at here is an integration of everybody who was from, say, 5 to zero, and 5 to 10, so that 10 bin window, 10-mm window on the basis of

baseline, and then it shifts over to 1 to 11, and then it's recomputed, and then it shifts over from 2 to 12, and so forth.

The good part about that is it gives you a smoothing. The bad part about it is it doesn't, unlike quintiles, provide any balance in sample size, so you end up out at the tail of this, you can see those three little dots at the bottom.

Those are all one subject. These are three points that represent one individual who started with a baseline score greater than 90, so it doesn't tell you anything about the density of the information and, in fact, some of these points on this plot have no information whatsoever in them.

So, whereas, over here, you might be looking at an integration over 40 or 50 patients, over here, you are looking at an integration over one patient who happens to have their value in that interval.

So, in the middle, it does a reasonable job, and, in fact, where you see where the quintile

lines are, it is reflecting things that are pretty similar to what we showed in the quintile analysis.

In the extremes, you can get wild reversals and all kinds of unstable estimates, and it doesn't give the impression of the lack of certainty in those estimates, whereas, the quintiles at least preserves the balancing of sample size. So, that is my technical definition of whacky.

DR. STOCKBRIDGE: Tell me what the basis was for doing quintiles instead of, I don't know, quartiles or deciles, or some other division.

DR. GIBBONS: A 1958 paper by William Cochran, a wonderful paper on stratification showing quintiles have wonderful mathematical properties, better than quartiles, and you don't need to go beyond quintiles, also forms the basis of most propensity score analyses.

DR. HIATT: I think that was a helpful clarification.

Overall, are the data compelling that patients with worse pain at baseline respond better

to Cellegesic?

I forget who gets to go here. Okay.

DR. LINCOFF: I think this shows exactly the same thing that was shown earlier. I mean it makes some physiologic sense. At the very highest range, you could say that the process is so severe that it is somewhat refractory, and like other therapies, in general, the worse you are, the more benefit you get. So, I think it is no additional information.

DR. HARRINGTON: Bob, can I ask you, does the same definition of compelling apply here?

DR. TEMPLE: Well, this isn't about evidence. This is about nuance and gut reactions.

DR. HARRINGTON: That was what I was hoping you were going to say.

DR. TEMPLE: I don't know if compelling applies to any of this.

DR. HARRINGTON: My read of this is similar to Mike's, that I think it is actually consistent with the previous analyses we have seen, and to me, this is a provocative

hypothesis-generating plot, that it may be that the only benefit is in a certain range of pain patients, and even that is modest.

DR. TEERLINK: I am not sure. I agree.

DR. HIATT: Dr. Koltun.

DR. KOLTUN: I don't understand the plot, but I think the more painful patients did the best is what it says.

DR. HIATT: Good. Thank you.

DR. TEMPLE: Whether you believe the middle part of that or the company's analysis on page 38 or whatever, it is suggesting like a 20-point move on that scale. That sounds fairly large if you believe that.

DR. HARRINGTON: That is exactly why I say I look at it as hypothesis-generating, because this is only one-fifth of an overall small population. Even if you were to plot the confidence intervals there, they would be broad.

So, for me, Bob, what this suggests is this is an interesting observation that may well be true, and would be an interesting follow-up study

to do.

DR. TEMPLE: One could look at the same analysis separately and see if it shows up each time, things like that. Okay.

DR. HIATT: I agree. I think it's consistent with the possibility of a subgroup, responder subgroup.

DR. GIBBONS: We did do it on all three studies, and it showed up in all three studies.

DR. WARNER-STEVENSON: I agree, but it is always easier to see a number fall when you start out high, but this is compelling.

DR. FINDLAY: I agree.

DR. KASKEL: I concur.

DR. PORTMAN: Yes.

DR. PICKERING: Yes, there is a subgroup that shows a favorable response.

DR. DeMETS: Suggesting and not compelling.

DR. FLACK: I agree.

DR. HIATT: We are coming up to the last question, and hopefully, the context is clear about

regulatory decision-making versus other kinds of clinical decision-making.

What is the appropriate regulatory action for Cellegesic? Please vote for one of the following options: approval, which our FDA colleagues would say you are done, approvable means that you are not, but the data are suggestive and more study might answer the questions that are not resolved, or not approvable, there is nothing compelling here, and you are done.

It's your vote, John.

DR. FLACK: I abstain.

DR. HIATT: You can't do that. This is where you have got to make a commitment.

DR. STOCKBRIDGE: This is why we pay you the big bucks to be here.

DR. HIATT: Exactly.

You have got to turn your mike on, too.

DR. FLACK: I am going to vote approvable pending another study.

DR. DeMETS: On the last key vote I said I had a very marginal yes for a lot of reasons. I am

not I guess sold on the outcome variable is one of my major problems. I don't know what it means especially. So, I would like to see something that is more concrete or definitive or something you can get your hands around, so because of all the other issues I had plus that one, I would vote approvable.

DR. PICKERING: I have been vacillating about this all day, but I think I am going to vote approval on the grounds that there is an unmet need, and it's a genuinely disabling condition. As I said earlier, I think there is some benefit and there appears to be a subgroup with more severe pain who do get a substantial benefit.

I am not too worried about the headache as a side effect because I think the patients will be told about headache and if they get it, they can decide whether the headache warrants discontinuation of the drug or whether the benefits that they are getting from the rectal pain may want them to continue with it.

I am not that concerned about the safety

issues, and I am not convinced that additional studies are really going to change the picture very much, because I think overall the effect size is quite small.

DR. PORTMAN: Vacillation time, no. Clearly, this is a tough one. There are things about the studies including the blood pressure issues that I mentioned before, and then maybe they can provide some other data here, but I think it is probably safe enough. I think there is a need for it although again I am concerned about the international flavor. Nonetheless, I think I will vote for approval.

I think that the headache is not that much of a concern. I think that people deserve the right to make a decision. If they have a significant amount of anal pain and we say, you know, this might cause you to have a significant headache, they say fine, bring on the headache, I want to get rid of this, and they deserve that right, and if the headache is too much, then, they can go see their colorectal surgeon, but I think I

would approve it.

DR. KASKEL: Despite the recent New England Journal article that I read on the plane about some of the problems with the subgroup analyses, and I shared with my colleague here, I think that I would vote to approve it.

I think that getting an additional study will not give much information more than what we have right now, and two, there will be further difficulties with patient enrollment, which I don't think we have addressed here.

The drugs are available, and I think people are just going to want to get this, and I think you will have trouble getting another study because of that.

DR. FINDLAY: I think clearly, further research is needed. I vote approvable pending another study of effectiveness.

DR. WARNER-STEVENSON: The biggest surprise to me today was the widespread use of this already, and I know that is not supposed to influence me, but I am afraid I can't forget it.

Nonetheless, I think I can put it aside for the moment and based on what we hear, and the difficulty of enrolling patients in any further trials, I would vote for approval.

DR. HIATT: My vote is approvable pending further study. I think that you just haven't convinced me that you have substantial evidence across two clean trials, and I am also impressed that although there may be a lot of off-label use, I don't think this is an impossible goal to achieve.

These aren't very big studies, these measurements are fairly simple, so they get to do a lot of kind of invasive testing to get your endpoint here. A comment came up you need 1,000 patients to prove this, would it be worth it, and we have discussed about that.

I don't think it would. I mean I think another properly powered trial maybe at 90 percent of targeting the population, you probably now have bracketed targeting the treatment window where you think the rate constant is linear.

I think all those things, you have that information now, and I think if you showed that, patients would know that treating themselves for a short period of time would perhaps give them some benefit, but today I am not convinced of that, and because of that uncertainty I would not vote for approval, but I think there is enough of a signal to warrant further study.

DR. HIATT: Dr. Koltun.

DR. KOLTUN: Approval.

DR. TEERLINK: I would vote approvable pending another study, as well, and reiterate much of Bill's comments. I think the challenge with dealing with single- symptom studies are difficult, and there are multiple issues that I have mentioned during the course of the day that need to be addressed before I would be more comfortable with approving.

DR. HARRINGTON: I also would vote approvable pending another study. I thought that the clinicians who presented today made a very compelling case that this is an important condition

that deserves better treatment and that this, in fact, might be one, but like my colleagues who have made the statement that you have not convinced us, you certainly have not convinced me.

I think that the evidence is not compelling, that the effect size probably does matter, and even if I put that aside, I am not sure what the effect size that you showed me means, and that the complete lack of any clinical correlate in the data has bothered me a bit throughout the review of the data and today.

So, I would say that while I appreciate the logistical challenges, I would vote for more research.

DR. LINCOFF: I previously concluded that I thought there was substantial evidence that there was some benefit. I also said that I didn't believe the magnitude of the benefit was important as long it's perceptible, and I think by definition in some way it had to be perceptible to be significant.

I think the safety issues are minor and I

agree that a patient should have the choice of having a headache rather than having rectal pain that feels like glass, and I also agree that I don't think much more evidence would be obtained from another study if it were performed, so I vote for approval.

DR. HIATT: Just to summarize the voting, we have just to be sure I have got this right: 6 for approval and 6 for approvable.

Statistically, we nailed that one.

[Laughter.]

DR. HIATT: Are there any concluding comments? If not, I would like to thank the sponsor and the FDA and all the committee members for some good work today. Look forward to seeing you all tomorrow.

We are adjourned.

[Whereupon, the proceedings were recessed at 5:25 p.m., to reconvene at 8:00 a.m., Wednesday, April 26, 2006.]

- - -