Considering Usual Medical Care in Clinical Trial Design: Scientific and Ethical Issues

November 14-15, 2005
Bethesda Marriott Hotel, Bethesda, Maryland

NIH Program on Clinical Research Policy Analysis and Coordination (CRpac)

NIH Program on Clinical Research Policy Analysis and Coordination

Considering Usual Medical Care in Clinical Trial Design: Scientific and Ethical Issues

Bethesda Marriott Hotel ♦ Bethesda, Maryland ♦ November 14-15, 2005

MEETING PROCEEDINGS

Conference organized by the
NIH Program on Clinical Research Policy Analysis and Coordination (CRpac)

Office of Science Policy
National Institutes of Health
U.S. Department of Health and Human Services

Acknowledgements

The NIH Program on Clinical Research Policy Analysis and Coordination (CRpac) would like to acknowledge the efforts of the federal interagency planning committee which during 2003 and 2004 developed the framework and plans for this meeting. Members of the planning committee included representatives of five federal agencies: AHRQ, NIH, OHRP, FDA, and CMS. Each of the agency representatives provided important input into the content and structure of the meeting. Each of the agencies, as well as representatives from the Centers for Disease Control and Prevention (CDC) and Department of Veteran's Affairs (VA), also participated in the federal roundtable session at the meeting.

We are grateful for the thoughtful presentations and discussion contributed by the speakers and panelists for this event. We gratefully acknowledge the drafting committee members who prepared a background briefing document in preparation for the meeting discussion. We also gratefully acknowledge all of the Clinical Research Policy Analysis and Coordination (CRpac) staff who worked diligently providing logistical and administrative support to the conference. A full listing of the planning committee, drafting committee, speakers, panelists, and participating CRpac staff are provided at the end of this report.

This meeting proceedings document was prepared by Donna Savage, science writer and contractor to the CRpac Program.

Table of Contents

Background

On November 14 and 15, 2006, the National Institutes of Health (NIH) sponsored a meeting to discuss clinical trial design issues that arise in the use of comparison groups that represent some form of current medical practice. In some cases, such comparison groups are easy to define because a standard regimen or treatment modality is well validated and widely used in the community. However, in many cases, uncertainty or lack of consensus exists about what kind of intervention should be used in a comparison group, or about how tightly controlled that comparison group should be. Often there is variability in clinical practice as well as disagreement about what practice to hold up as the best standard. Disagreements also arise about whether current practice is supported adequately by data from previous clinical trials.

This variability and lack of consensus can pose a particular challenge for clinical trialists seeking to rigorously test new or existing interventions in a trial. Should usual medical practice be used as a comparison group, even when that practice has not been validated scientifically? Should each specific intervention used in the community setting be studied as a separate and discrete arm in a clinical trial, or should a range of practices be allowed within one trial arm? These kinds of questions formed the basis for the discussion about the complexities of considering usual medical practice in designing clinical trials. The meeting was attended by more than 300 people from government agencies, academic medical centers, and industry.

The morning session on the first day consisted of a number of didactic talks on basic topics: variation in medical practice, clinical trial design, ethics of randomized controlled trials, and the practice of evidence-based medicine. The afternoon session and the morning of the second day were devoted to case studies of actual clinical trials illustrating different choices of comparator arms in situations in which usual medical practice was complex or variable.

Three case studies were selected in different areas of medicine: oncology, mental health, and surgery. The first case involved a trial of treatment for ovarian cancer. The trial, ICON-5, was particularly interesting in that it represented one of a series of trials by collaborative groups in the United States, the United Kingdom, and Canada, and each trial built upon previous results from that consortium as well as results from other clinical trials. Differing interpretations of previous trial data led investigators to harbor differing views about the most appropriate control arm.

The second case study was a multisite study of attention deficit hyperactivity disorder (ADHD) treatments – the Multimodal Treatment Study of ADHD, or MTA. This study has been recognized as an important milestone in clinical research on the topic of ADHD treatment due to its size, careful design, and consideration of both behavioral and pharmacological best practices. The study included a community comparison group that consisted of usual care delivered in the community setting. The main

question being addressed in the trial was whether behavioral treatment, drug treatment, or a combination of the two would be most effective in treating children diagnosed with ADHD according to standard criteria. The behavioral and pharmacologic interventions were structured as "gold standard" best practices, which differed, in many cases, from the kind of care offered in the community setting. Therefore, it was of interest to consider what the community comparison group added to the trial in terms of interpretation of results and generalizability of findings.

The third case study described an ambitious and innovative trial of surgical and nonsurgical treatments for back pain – the Spine Outcomes Research Trial, or SPORT. This trial has enrolled patients in either an observational or controlled trial prior to treatment. Those patients who decline to join the randomized controlled trial are asked if they would enroll in an observational study that will collect data on treatment choice and outcomes; treatment is chosen by patients and their doctors as usual. Those who choose to enroll in the controlled trial will be randomized to nonsurgical or surgical treatment. The nonsurgical treatment arm allows patients to choose among an extensive list of nonsurgical alternatives. The observational component of the study allows investigators to learn about the characteristics of patients who decline randomization, perhaps because they have strong treatment preferences. The flexible nonsurgical comparison group allows patient choice, while at the same time collecting data about whether surgical or nonsurgical treatments are more effective for this group of patients.

The meeting concluded with a roundtable discussion of the clinical trial design issues that had been explored at the meeting, with the goal of developing some general recommendations and approaches to this topic. Discussion centered on several basic questions:

a) How can one determine whether a flexible usual care arm is appropriate and informative in a given setting? Are there some general rules that can be developed to help with these decisions?

b) When a usual care arm is included in a trial, what particular ethical issues may arise, for example, in relation to the possibility of substandard care that may exist in the community? What kind of precautions need to be taken to protect the welfare of research subjects, and how far do researcher obligations go in terms of intervening in the practice of usual care in the community?

c) How can results of trials with heterogeneous comparison groups, such as usual care groups, be interpreted? How do trials including these comparison groups fit into a longer sequence of clinical trials and accumulating evidence on a given clinical research topic?

No simple approach to these complex issues was uncovered, but some important scientific and ethical concerns were raised. These concerns will be addressed further in a "points to consider" document that will elaborate on the considerations that

investigators should take into account in weighing the need for a usual care arm when designing clinical trials.

Welcome

Raynard S. Kington, M.D., Ph.D., Deputy Director, National Institutes of Health

> *Dr. Raynard S. Kington was appointed Deputy Director of the National Institutes of Health (NIH) on February 9, 2003. Prior to coming to NIH, he was Director of the Division of Health Examination Statistics at the National Center for Health Statistics (NCHS) of the U.S. Centers for Disease Control and Prevention. As Division Director, Dr. Kington also served as Director of the National Health and Nutrition Examination Survey, one of the Nation's largest studies to assess the health of the American people. His research has focused on the role of social factors, especially socioeconomic status, as determinants of health among elderly persons, and racial and ethnic differences in the use of long-term care.*

On behalf of the NIH, it is my pleasure to welcome participants to this conference on Considering Usual Medical Care in Clinical Trial Design: Scientific and Ethical Issues.

This event will deal with a critically important issue: in designing clinical trials, how we maximize the scientific and ethical integrity of research. Appropriate clinical trial design is vital to ensuring scientific validity, to providing societal benefits from knowledge gained, and to upholding ethical principles in human subjects research. The specific research questions being asked in clinical trials shape this design. Often, choices about the specific parameters of trials are not straightforward.

You will address just one area in depth during the next two days: How "usual care" should factor in to the design of comparison arms of various trials. In some cases, there is no clear consensus on which comparison arms should be included in a clinical trial testing a new or existing intervention. The classic approach in many trials is to compare an intervention of interest to standard treatment, but it can be difficult to determine a uniform standard treatment in some cases, especially when multiple treatment modalities are used or when there is a lack of consensus in the professional community regarding which treatment should be considered the best standard. When more than one treatment is available, there may also be disagreement about the criteria for selecting a specific treatment for individual patients.

The purpose of this meeting is to have a broad dialogue about the scientific and ethical dimensions of this issue. Equally important, we hope to lay down the intellectual foundation for a "points to consider" document to offer information and guidance to the research community on this topic.

I cannot think of a time when such issues have been more important to the clinical research enterprise. Some of you may have seen the front-page lead article in this morning's *New York Times*, which suggested that erosion in public trust about pharmaceutical company research is now beginning to have an impact on the sales and profits of these companies, which work is important in the commercialization of scientific advances into products available to the public. Efforts such as this conference to ensure the highest ethical and scientific standards are essential to building public trust in our clinical research enterprise.

This conference was planned under the auspices of the NIH Clinical Research Policy Analysis and Coordination (CRpac) program. The CRpac program is an important element of one of the initiatives of the NIH Roadmap, which is NIH Director Dr. Zerhouni's plan to optimize the ability of the Nation's research enterprise to tap into the resources of the country, advancing the research enterprise and facilitating the translation of scientific findings into real advances in clinical practice. In keeping with this global aim of all Roadmap initiatives, an objective of the CRpac program is to bring Federal Agencies, academic partners, and private organizations together to work toward productive solutions to some of the significant challenges being faced by the clinical research community. With participation and input from across the Government and the private sector, this meeting is one way in which the CRpac program is working on an array of clinical research issues.

Before I turn the podium over to Dr. Amy Patterson, the director of the NIH CRpac program, I want to thank everyone for being here today. We are fortunate to have some of the top experts in the Nation and abroad with us today to explore these issues. We appreciate the efforts of the speakers, the panelists, and the audience members who will be invited to join in the discussion of this complex issue. Your thoughtful participation in this meeting will help us achieve the most useful outcome possible. The time you are taking here today is much appreciated and we hope there will be a concrete product coming from this discussion that will help other researchers from across the country and around the world deal with these issues.

Welcome and Introduction

Amy P. Patterson, M.D., Director, CRpac Program, Office of Science Policy, NIH

> *Dr. Patterson is Director of the NIH Program on Clinical Research Policy Analysis and Coordination (CRpac) and Director of the Office of Biotechnology Activities (OBA), both within the Office of Science Policy, Office of the Director, National Institutes of Health (NIH). In the first capacity, she leads a program which provides a focal point for streamlining, coordinating, and harmonizing Federal policies concerning the conduct and oversight of clinical research. In directing the OBA, Dr. Patterson oversees the management of several programs concerned with science, safety, and ethics in a number of critical fields of biomedical research. In addition to her management responsibilities as Director of OBA and the CRpac Program, she is on the clinical staff at the NIH Clinical Center and maintains an active basic research program at the National Heart, Lung, and Blood Institute.*

It is my pleasure to echo Dr. Kington's welcome and extend a warm and appreciative welcome to our esteemed speakers and panelists as well as to the diverse audience. The fact that we had to close registration more than a month ago is a testament to the research community's interest in this topic – how to incorporate considerations about usual medical care into the optimal design of clinical trials. It is our hope that, in the next 2 days of dialogue, sharing of experiences, and working through the case studies, we will move quickly from the conceptual to the practical and emerge with the beginning of a conceptual framework that will guide intelligent and thoughtful thinking about the scientific and ethical design of trials.

The "Points To Consider" document being used as a springboard for this conference is purely a starting point to launch the dialogue; it is very much "written in pencil." We hope that, at the end of the 2 days, the dialogues that go on here will help refine the principles and that we will emerge with the beginning of an appropriate conceptual framework that will be further vetted through the research community and through our sibling Federal Agencies, all of whom have a keen interest in this topic.

I want to acknowledge the interest and input from our sibling Agencies: FDA, AHRQ, CMS, VA, and OHRP as well as from our own "family" – the multiple Institutes and Centers of the NIH that have provided input into the design of the conference.

Introduction to the Conference

Robert J. Levine, M.D.
Yale University

Introduction to the Conference

Robert J. Levine, M.D., Yale University

Dr. Levine, meeting chair, is professor of medicine and lecturer in pharmacology at Yale University School of Medicine. He directs the Law, Policy, & Ethics core of Yale University's Center for Interdisciplinary Research on AIDS, and is co-director of the Yale University Interdisciplinary Bioethics Center. Dr. Levine is also author of a leading text entitled, "Ethics and Regulation of Clinical Research" and he has leant his experience to matters regarding the ethics of research involving human subjects to the Council of International Organizations of Medical Sciences, the World Medical Association, the Joint United Nations Program on HIV/AIDS, the Pan American Health Organization's International Bioethics Advisory Board, and various institutes of NIH and CDC – among many other organizations. Dr. Levine chaired the Yale Institutional Review Board (IRB) for 30 years and served as special consultant to the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research as it developed the landmark document, "The Belmont Report."

What are the objectives of this conference? In this conference we will consider the circumstances in which usual care, or some subset of usual care, should be used as a comparator in clinical trials. Organizers of this conference hope to see three specific outcomes:

- Develop a proceedings document to serve as valuable resource for those interested in the conduct of clinical trials.

- Develop a points to consider document, the draft of which you already have. This draft will be the subject of discussion tomorrow afternoon at the roundtable.

- Devote attention to process issues – e.g., how do we know we have invited all of the relevant stakeholders to participate in the discussion leading to the design of clinical trials, and how to collect and evaluate the best evidence.

The idea for holding this conference was born in the discussions of the ARDSnet trial. There were criticisms of various aspects of the trial, particularly whether there should have been a "usual care" control arm. The specific request to convene this conference came from the Office for Human Research Protections (OHRP). The ARDSnet trial will be a subject of discussion this afternoon. It is not our purpose to second-guess any of the decisions made in that case; they have been amply and extensively discussed elsewhere. Our purpose is to consider this trial in order to discern lessons learned from this experience.

We will be concerned only with trials involving human subjects that involve prospective assignment to an intervention, that have two or more comparison arms, and that have health-related outcomes. This means we are excluding retrospective studies, uncontrolled trials, and prospective observational trials. We will focus on trials that are designed to be relevant to clinical practice rather than those that are designed to measure basic biological processes or disease mechanisms. Many trials include objectives in both of these categories; of these, we will consider those trials in which the

primary outcome measures are directly relevant to clinical practice and in which secondary outcome measures are directed at basic mechanisms and processes.

Some important topics are not the main focus of this conference, but because they are relevant they will probably be mentioned during the discussions. These include placebo-controlled clinical trials, clinical trials in developing countries, the role of funding agencies in shaping clinical trials, research in quality improvement, and evaluation of clinical equipoise as an ethical construct. (In a short conference such as this, one cannot highlight all the important issues.)

A major event in this conference will be the presentation and discussion of three case studies. First we will discuss ICON-5, the ovarian cancer treatment trial; the main reason this trial was selected for discussion is that it was conducted in both the United States and the United Kingdom, each with differing standards of care. Secondly, we will discus the MTA trial, which is a multi-modal treatment study of attention deficit and hyperactivity disorder (ADHD). This trial was a comparison of drug treatment versus behavioral treatment. The main reason it was included in the discussion was that it contained a community comparison arm that allowed anything the physicians ordinarily did in their regular practices. The final discussion will be of the spine patient outcomes research trial (SPORT), in which the comparison is between surgical and nonsurgical approaches to treatment. One arm of this trial is the patient's choice, which is the main reason for discussing this trial.

Some topics might come up that are important but are not targeted for discussion. These topics include the issue of "super-care," particularly in medical care comparison arms in trials of surgical interventions. Often in these trials, the medical therapy provided for the so-called control arm of the trial seems, to those responsible for the trial, to be far superior to what is provided in ordinary medical practice; we wonder whether this erects too great a barrier to establishing surgical intervention as safe and effective. Another topic that might come up is what to do when the "approved" drug for the condition is one that practitioners do not use. In addition is the concept of the "ethical imperative" of "the good of personal care," which was introduced in the 1970s by Charles Fried and more recently has been discussed primarily as the fiduciary obligation of the physician-investigator.

Practice Variations: Some Implications for the Design of Clinical Trials

John E. Wennberg, M.D., M.P.H.
Dartmouth Medical School

Practice Variations: Some Implications for the Design of Clinical Trials

John E. Wennberg, M.D., M.P.H., Dartmouth Medical School

Dr. Wennberg is director of the Center for Evaluative Clinical Sciences at the Dartmouth Medical School and professor in the Department of Community and Family Medicine and the Department of Medicine. He became interested in the application of epidemiological principles to the healthcare system while pursuing his master's degree in public health at Johns Hopkins. With colleagues he has developed a strategy for analyzing the population-based rates of health resource allocation and utilization. He has made contributions to the importance of patient preference in the rational choice of treatment.

My research has been to look at the way healthcare is delivered from one geographic area to another, called small area analysis. More recently, we have been looking at variation between specific provider groups and, it turns out, there is as much variation within a region as there is between them. They all have relationships and importance to the design of clinical trials in which usual practice is part of the arm. The question is: what is usual practice? It turns out there is no such thing! I will take you through some examples of this practice variation within a framework we have been using – that variation occurs according to three categories of medical care:

- Effective care that clinical trials have shown to be useful and for which there is no tradeoff; e.g., a beta-blocker after a heart attack, in which the problem is under-use of effective care. That matters in clinical trials because, if you are randomizing people and your control arm is not compliant with practice guidelines for beta-blockers, you will get confounding effects on that basis. That is an obvious problem on which I will not spend much time.
- Preference-sensitive conditions
- Supply-sensitive conditions

A paper that Sean Tunis, Caroline Clancy, and Dan Stryer wrote some time ago discusses the clinical trial problem in terms of practical clinical trials or trials that are trying to improve decision-making in clinical or health policy. The practice variations we are looking at compare practices around the United States using Medicare data (the only national database we have) to characterize variability within local communities.

I am starting out with the preference-sensitive treatments, the obvious case of which is discretionary surgery. One of the most classic examples would be whether a woman with breast cancer is treated with a lumpectomy or a mastectomy. Clinical trial evidence shows these treatment modalities to be equivalent in terms of survivorship but all other outcomes are different. Normatively speaking, it seems clear that that decision belongs ultimately to the patient because it is a trade-off among several unpleasant things and people have different ways of evaluating the degrees of unpleasantness. These tradeoffs involve more than one treatment and different outcomes. Clinical evidence is sometimes good, as in the case of breast cancer, and sometimes clinical evidence is not so good or is in dispute, as in the case of prostate cancer. The decision should be based on the patient's preference, but the evidence from observational interpretations and clinical trials is that provider opinion will often determine which treatment is used.

Here are some examples to show how much care varies, comparing the region immediately around Stanford University Hospital (the only hospital in that region) to San Francisco (which has several hospitals) and Los Angeles. Stanford excels in back surgery, doing 2.2 times as many procedures during a given period of time than do physicians in Los Angeles or San Francisco. In terms of knee replacement, it's a tie – no one is doing a lot of them. But if you look around the country you will see a four-fold variation in the rates of knee replacement. For back surgery, Stanford has a really aggressive group looking for that. Back pain is not more prevalent in Stanford than it is in Los Angeles or San Francisco, but there is a predilection on the part of providers in those communities to concentrate on a particular procedure.

We find indirect evidence of the relevance of provider opinion in the correlation of the supply of orthopedic surgeons in each of these regions with the rates of knee replacement. If I asked you before I showed you the data, most of you would think there would be a correlation between the supply – the capacity of the system to provide orthopedic surgery – and the numbers of orthopedic surgeons actively engaged in doing knee replacements. It turns out that is not the case. What happens behaviorally is that individuals or small groups become expert in one specific subset of surgical activity and concentrate on it. Surgeons with orthopedic training can do carpal tunnel, sports medicine, knees, hips, backs, or trauma. Specialization is responsible for this non-correlation.

This is not true just for orthopedic surgery; it does not matter which procedure we are looking at – there is not a strong association between supply and utilization rates for these kinds of decisions, but there is a strong relationship over time with procedures. In this graph, on the horizontal axis we have knee replacements at the beginning of the 1990s correlated with knee replacements at the beginning of the 21st century, and you can see by the location of the black line that each region does more than it did before but we do not see any evidence of regression to the mean. This is a fixed attribute of the community in which we are practicing, which will be true of most surgical procedures. The R-squares between the beginning and the end of the 10-year observation period are all significantly higher.

The last decade has seen the evolution of a number of patient decision aids that lay out the known costs, risks, and benefits of these procedures in a standardized format, and then are used in clinical trials to find out the effect of decision aids on treatment choices. We always see a difference between the control arm and the experimental arm, even though the control arm is hardly unbiased because it is usually the same patients for the same physician participating in the clinical trial.

What we see is not only that decisions are different in the experimental versus control arms, but also the evidence that they are better decisions. What do I mean by that? Here is a study done in the early 1990s that was designed to provide relevant information about benign prostatic hyperplasia choices and consequences on clinical uncertainties. The study also had methods of measuring the patients' preferences with regard to the tradeoffs involved in this decision. Once these decision aids were in place, BPH surgery dropped 40 percent in these two staff-model HMOs in which this

study was being conducted, a benchmark that was close to the bottom of the national distribution. This says that, at least from that benchmark, the amount of surgery being delivered in the United States for this condition at that time seemed to exceed in most regions the amount that well informed patients would want under this paradigm of evaluation strategy.

The evidence that the decisions were better came from two sources. Once was the fact that the control arm was less informed and was not able to describe what was at stake, so the knowledge base was different. Secondly, the concordance between the values of the patient and the decision made actually improved. While symptom level was the reason for operating, for patients who were severely symptomatic and who were offered BPH surgery, only about 20 percent chose it. In other words, the clinical evidence that this was an appropriate intervention, with which all urologists would agree, failed to take into account patients' different points of view. What we saw in multivariate predictors was that what really mattered was not your symptom score but your attitude toward symptoms – how much did they bother you and (for BPH) your concern about impotence and retrograde ejaculation. It was a tradeoff between sexual function and urinary tract function, and men differed on that, not surprisingly.

When considering the implications for design of clinical trials, one is immediately confronted with two fundamental problems – the problem of patient preference for the outcomes and questions about the probability of those outcomes occurring. The clinical trial "industry" has concentrated almost exclusively on probability rather than patient preference issues. This will emerge more importantly as a role that needs to be looked at under the clinical trial umbrella. Dr. Weinstein will argue tomorrow that the standard of equipoise in these preference-sensitive conditions should presumably be the patient's point of view and not the clinician's point of view. Equipoise according to patient preference may lead to different trials, trial designs, and uptakes compared with the classic concept. This entails a shift from informed consent to informed patient choice. This juxtaposes the classic equipoise, between two providers who disagree about the evidence, to the question of how the patient processes information about clinical uncertainty – and how that affects participation in clinical trials.

The third category is clinical decision-making for supply-sensitive care, which involves management of chronically ill patients: how frequently they are visited by a physician, how many times they are hospitalized, how many times they enter intensive care units, and how many times they have imaging exams. As I will show, this varies significantly from one place to another. Under the current regime of clinical uncertainty, the use of these treatments is governed primarily by the assumption that resources should be fully utilized – doctors always fill up their offices and hospitals try to fill up their beds. The assumption is that more is better. But medical theories and medical evidence play almost no role in governing the frequency of use. You cannot go to any practice guidelines book and find a reference to how frequently you should visit a patient at any given level of illness, when to hospitalize, or when to put a patient in intensive care units. These topics are not discussed at clinical rounds and have probably never been raised at the NIH. Yet it turns fundamentally on the variation in cost between Medicare regions. The reason that Miami spends twice as much per capita as Minneapolis maps

directly to how frequently chronically ill patients are put in the hospital and seen by their doctors.

The observational evidence for the association between capacity and discharge rate shows virtually no relationship between bed supply and the hospitalization rate for hip fracture. Patients with hip fracture almost always seek care and almost always get diagnosed, and no physician in the United States would say to treat them as an outpatient. The illness rates drive the utilization rates, and they are independent of bed supply. But that is not true for almost all other medical conditions – for example, congestive heart failure, COPD, or any cancer. The same goes for the number of cardiologists and the number of visits. Cardiologists do not have unfilled office hours; cardiologists in high supply zones fill up those hours by shortening the interval between revisits or with more referrals.

The obvious question here is: so what? It would be a shame if these specialists were not working, but the question is whether more is better – is it improved health care? There is not a shred of experimental evidence that would allow us to deal effectively with that question. There are, however, observational studies that indicate clearly that there is no marginal gain in life expectancy, qualify of life, or quality of care as one moves across the gradient from Minneapolis to Sacramento to Miami. There is no evidence that those people in "super care" regions will benefit.

More recently, we have been looking at hospital-specific measures. A two-year follow-back study from death included 12 chronic illnesses accounts for about 75 percent of Medicare deaths. People who have these diseases tend to be hospitalized. If you assign people to the hospital they most frequently used in the last two years of life, you end up with a population of chronically ill patients all of whom have similar prognoses. The argument about whether they were sicker in Los Angeles than in Minneapolis comes down to the question of whether you are "deader" in one place than another; so far we have not had much debate about that! This mitigates one of the critical shortcomings of observational studies by this methodology.

In this process, we have looked across the entire United States. Although some of our data will not be out until early January 2006, I want to show some of the things we are finding for academic medical centers, which will give you a sense of the importance of differences in usual care patterns for any kind of clinical trial design around the treatment of chronic illness.

This graph shows the number of days in the intensive care unit in the last 6 months of life among Medicare enrollees receiving most of their care at one of the 120 Council of Teaching (COT) hospitals, which are integrated with medical schools. In some places, we see less than two days per person in the last 6 months of life and in others up to 11 days per person in the last 6 months of life, showing very striking differences in what "super-care" is all about, depending on the academic medical center. At the University of California system, days in intensive care go all the way from 3.3 days at UCSF up to 11.4 days at UCLA, so these places are practicing very different kinds of care in

managing chronic illness. There are huge implications of the costliness of managing chronic illness between these different institutions.

Physician visits go from less than 20 up to 77. Across the spectrum in California, you see everything from UC Davis and Stanford at the bottom up to UCLA doing 52 visits. The champion of super-care is NYU, which is not known for taking care of sick people who are minorities, so it is not explained by the usual concepts.

This chart shows the correlation among these academic medical centers between cancer patients on the horizontal axis and congestive heart failure patients on the vertical axis. What really matters here is the institution in which you are getting care, not the illness you have. It is true for physician visits and it is true for hospital days. What we are dealing with here is a systems attribute that is affecting clinical practice in a subliminal way, of which the participants are not fully aware, which flies in the face of our belief that all medical activity is purposeful.

We see the same thing across time. This chart looks at the severity of illness, indirectly measured. Along the horizontal axis we see the same people but 19 to 24 months before they died compared to their activity in the last 6 months of life. You can see that you get a lot less care in the 19 to 24 months prior to death. This attribute is fixed and it affects not just end-of-life care; it has to do with how chronically ill people are managed over time in these different institutions.

How many doctors get involved in your care is also an indirect measure of super-care. This graph shows the percentage of patients at these academic medical centers who see 10 or more physicians in the last 6 months of life: from less than 20 percent to more than 60 percent. Within the California system, at UCLA you get shuffled to all sorts of doctors, quite literally, and in the northern part of the state much less so.

The nature of the workforce is interesting. Some academic medical centers depend much more on medical specialists than they do on primary care physicians (internists and family practitioners). This chart expresses the ratio of visits to medical specialists (on the top) to primary care visits. Some places are seeing 2.5 times more medical specialists than they are primary care physicians; others are seeing more primary care physicians than they are medical specialists. I am particularly interested in the difference between UCSF, which is primary care oriented and where the ratio is less than 1, and UCLA where the ratio is 2.8.

Here is a little evidence about what happens if you try to design your observational studies within your own systems. I want to look at differences between hospitals that belong to the Mayo Foundation system of care. The system average is 2.8 days in intensive care unit (ICU) but the range is from 7 days to almost none in some of the smaller hospitals. Of most interest in this diagram would be what is going on at St. Mary's Rochester Methodist, at Scottsdale, and at Jackson. They differ extraordinarily, even within Rochester, Minnesota, itself. The practice variation issue is deeply embedded into the matrix of clinical practice, even when managerial sharing occurs across institutions.

In contrast, Fairview Health in Minneapolis (which includes the University of Minnesota) has one of the tightest distributions; it would be interesting to find out what is going on there. The Cleveland Clinic has, 4.6 days on average in ICU but there are big differences with this system. This chart compares university hospitals, showing wide variability. If one were to randomize on the basis of hospitals to do a clinical trial in which utilization and cost were some of the outcomes, the background "noise" in this system would likely overwhelm most of the effects associated with the intervention.

The Care Group in Boston compared to the Partners Healthcare at Massachusetts General and the Brigham are quite different in terms of this particular parameter, as are some of the associated hospitals. On these others, the pattern of physician visits is the same – big variability – as is the ratio of medical specialists; everyone using is their workforce in different ways.

For practical clinical trials, we need to concentrate on understanding the relative importance of preferences versus prognosis, and we need to design our clinical trials to take those into account. Particularly, we will flounder in surgical trials on the problem of the placebo effect, which is an almost unsolvable problem.

In terms of randomized clinical trials (RCTs) in which supply is important, here are three trials in which the systems effect was not taken into account:

- The Rand Health Insurance Study randomized only a small proportion of people within each market area. The assumption was that if the effects that were observed were generalized, that would represent a reasonable estimate of the effects of co-payments on the system. The system was not challenged by this small randomization; it was at the margins. Therefore, the impact at the individual level probably does not predict the effect of system-wide changes in co-payment that would, in effect, require downsizing of capacity and otherwise dealing with the fundamental supply dynamics that are behind the variation problem.

- The SUPPORT Trial of End of Life Care was a large trial trying to find out whether advance directives would reduce utilization. It had no effect at all. We re-analyzed the SUPPORT trial data to show that the utilization rates of those communities were correlated with the capacity of the local system. The advance directives had no effect on the fundamental dynamics that were forcing the system to behave the way it did.

- Section 731 of the Medicare Modernization Act (MMA) trial has spawned chronic disease management RCTs. These trials are the subsets of patients with congestive heart failure; the control group is people in the same region with congestive heart failure. There may well be an effect associated with chronic disease management in that trial, but the overall outcome is the cost of payment. The value of chronic disease management will be lost without taking into consideration the fundamental behavior associated with the fact that management of all forms of chronic illness, including congestive heart failure, are

correlated. If you reduce one, it is likely you will bump up the rest of them and therefore get a false sense, if you are looking only at the experimental group, of the impact on costs.

Increasing the value of clinical research for clinical decisionmaking and health policy requires a broad approach that:

- Takes local practice patterns and supply into account in designing systems and interpreting results
- Addresses preferences (values) as well as prognosis (probabilities)
- Mobilizes appropriate disciplines and uses a wide range of study methodologies

The artificial constraint on me at this conference is that I cannot talk about things that matter most, in terms of moving the theory to the table where clinical trials might actually be relevant. So much can be learned by observational studies. In our work with prostate disease, we clarified the medical theory that it was not associated with chronic obstruction but was a symptom level decision, based on preferences. Surgery had a slam-bang effect, but the fundamental problem was preference – what do men want, under what circumstances do they want it, and how does it impact their lives.

Discussion

Q: Please elaborate on your comment on Section 731.

Dr. Wennberg: There is a trial that focuses on chronically ill patients with congestive heart failure and diabetes. The quite-reasonable design was to have a control group of individuals not in the chronic disease management program and a group of people that are in the program. We will sample them in the population of the region. The underlying query is how chronic illness management influences overall utilization and overall costs of care – and presumably outcomes.

The difficulty with that is, because of the role of the invisible factor of capacity in determining how much services are used, if you decrease the use of one component of care in a market through an intervention, and your control group is not using that care, it is likely that the cost effectiveness of the control group will go up and that of the experimental group will go down. However, you are really interested in the overall impact on the system as a whole. The rule of supply and supply-sensitive services is that if you cannot get a service for X then you will get it for other causes. The implication that you would get the same result if you fully mobilized the system is not predicted by the results when you are just asking about the two arms of that trial.

Q: You said that usual care does not exist and cannot be defined. You also emphasized mitigating the system attributes and described how important it was to articulate them and address them in trials. What role might explicit methodologies for experimentation have on mollifying the influence of these effects on the overall system response in the clinical trial? In other words, providing specific decision support rules

for those engaged in the trial in order to try to make more uniform the behavioral attributes of the centers enrolling patients.

Dr Wennberg: That would certainly standardize things, but what generalization can be made from that? Clinical practice is so heterogeneous that, if you artificially constrain it in the control arm, you are actually doing two experiments at once.

Q: That would depend on what tools you use to effect the explicit methodologies, where available, for future use in practice and you had a method for making it happen.

Dr. Wennberg: Then it would not be a control group; you would have two experimental groups.

Dr. Levine: It is important that you showed us that usual care does not exist. There are parallel considerations in the development of guidelines. In the late 1990s, the National Bioethics Advisory Commission was asked to develop standards for clinical trials conducted in other countries. It departed from what was then the prevailing standard, in such things as the Declaration of Helsinki, which referred to the best-proven therapeutic method and recognizing the multiplicity of treatment methods for almost everything. This changed the standard in international documents to establish effective intervention, recognizing that there would be quite a number of them for each particular condition. This cue was picked up by the Council for International Organizations of Medical Sciences (CIOMS) and became the standard in that document as well. But how would you conduct a multicenter trial with 60 or 80 centers?

Dr. Wennberg: That is the way out of our dilemma, to chip away at it.

Dr. Haynes: If there is not a good correlation between the amount of care provided and patient outcome, then the variation across communities is not that important when conducting randomized trials across sites. Secondly, if you conduct trials for people who do choose to take an intervention, then at least it would be a result that should be generalizable to people who would choose that option. Of course there will be people who will not take that option, but that is not relevant to the trial because they will not choose it anyway. So what is the point about the implications for randomized trials?

Dr. Wennberg: For people who are in randomized trials, the data only represents what an actively choosing patient would get. Whether or not it matters to a particular trial whether you are conducting it in Miami or in Minneapolis or within these academic medical center chains probably would depend on what you are trying to infer from the results. If you are trying to infer that intervention X (e.g., in the CMS trial of chronic disease management) has a system effect as opposed to an individual patient effect, you need to be very careful about that. What drives utilization overall is not scientific algorithms but rather the available capacity, in that example. It is a point that needs to be taken into account seriously. What this means is that we need to pay a lot more attention to what this all means in terms of patient outcomes and value. If it does not produce anything different but it costs twice as much, that identifies a great deal of waste in the system that needs to be dealt with.

Q: An article recently in the *New York Times* talked about a study about choosing between lumpectomy and mastectomy. When given the choice, patients chose mastectomy and doctors chose lumpectomy. (The article did not mention geographic variations.) This seems counter-intuitive.

Dr. Wennberg: The clinical trials of the decision aids have shown for this condition that some women are very averse to the idea of local recurrence and dislike chemotherapy, and some want to preserve their breasts so choose lumpectomy. There is no way to diagnose that without asking women what they want on an individual basis. That article missed the important point – that this was a preference-based choice and it depended on patient preference and not doctor preference. That was implied in that article but never quite accurately teased out.

Dr. Levine: That was a complete turnaround from what the prevailing presumptions when the debate came up in the early 1980s, when it was assumed that all the doctors would want mastectomy and all the women would want lumpectomy.

Dr. Wennberg: This idea of believing that there is a single right answer for a condition that has tradeoffs is what we are trying to get on the table. It is an assumption that, for example, if you do a clinical trial of a blood-sugar-lowering agent or find out that A1C hemoglobin really works well, some people find the regime to get that so abhorrent that their preferences run in a completely opposite direction. As a society we do not deal well with information exchange and letting patients have a choice in their treatments, yet that is so implied in all the ethical arguments. It is behind our interests in shifting the legal standard of practice from informed consent to informed patient choice when dealing with these preference-sensitive conditions. The language gets entangled in this antiquated system of making clinical decisions.

Q: Your research tends to deflate the claims that defensive medicine accounts for a general increase in utilization of care. You are showing places within the same legal jurisdictions that indicate a dramatic difference in practice patterns. Therefore, claims that defensive medicine is the cause of super care tend to be deflated by this research. Do you have any evidence about what happens outside of academic medical centers, in terms of the data?

Dr. Wennberg: Next week we will have a paper showing variations in California, where the observations are focused on one region (e.g., Sacramento versus Los Angeles) and then, within regions, variation in all comers. In terms of both high use rates and costs-per-person, the academic medical centers are not the leaders in Los Angeles; some small private hospitals are even more aggressive in managing this kind of stuff. This phenomenon is available in each region and each system.

Dr. Levine: We have begun to use the term “equipoise” in this discussion. Reference was made to equipoise from the patient point of view. It is important to note that the initial introduction of equipoise as part of the concept of clinical equipoise was by the late Benjamin Freedman: he attached the idea of clinical equipoise to a position held by the expert clinical community. It is important to recognize that all of the ethical

conclusions from the discussion of this concept, defined as he did, are dependent on that definition. If we turn to another use of equipoise that means we cannot draw the conclusions of Freedman or his followers. When we use a different sense of the term, the old definitions no longer apply.

Dr. Wennberg: In Dr. Weinstein's clinical trial, which he will talk about tomorrow, patients it was fully explained to patients what was known and not known about the treatment options. They were then asked to enroll in the clinical trial; followup included those who agreed to be randomized and those who did not. Whether we used the right term or not, the ethical basis of that trial was that some fully informed patients chose to be randomized, for whatever purpose. Secondly, a large proportion of them did choose to be randomized, relative to clinical trials governed by the old idea that the doctors had to decide who was a candidate for randomization under the Freedman doctrine. I am not a scholar in this and I do not mean to upset the vocabulary, but we need to make a strong distinction. Surgeons do not like to disagree. The concept is that there is lack of evidence that would lead the physician to say, "this is what is not known" and then that information is conveyed to patients, and patients actively choose. The language needs to be worked on, because ultimately we will not get far unless we have a common vocabulary.

Dr. Levine: I do not mean to suggest you are misusing language. The roots of this term in the discourse in ethics depend upon Freedman's initial definition. It was clear to everyone what you meant when you talked about equipoise from the patient's point of view.

Federal Agency Views on the Relevance of Usual Care Control Groups

David Atkins, M.D., M.P.H.
Agency for Healthcare Research and Quality (AHRQ)
Louis Fiore, M.D., M.P.H.
Department of Veterans Affairs (VA)
Chad M. Heilig, Ph.D.
Centers for Disease Control and Prevention (CDC)
Amy P. Patterson, M.D.
National Institutes of Health (NIH)
Marcel E. Salive, M.D., M.P.H.
Centers of Medicare and Medicaid Services (CMS)
Bernard A. Schwetz, D.V.M., Ph.D.
Office of Human Research Protection (OHRP)
Robert J. Meyer, M.D.
Food and Drug Administration (FDA)

Federal Agency Views on the Relevance of Usual Care Control Groups (FDA, NIH, AHRQ, OHRP, CMS)

David Atkins, M.D., M.P.H., Agency for Healthcare Research and Quality (AHRQ)

> *Dr. Atkins is chief medical officer in the Center for Outcomes and Evidence at the AHRQ. In that role, he is responsible for coordinating the Agency's research portfolio on care management and providing scientific oversight of the work of the 13 AHRQ evidence-based practice centers. He was a student member of the IRB at Yale Medical School, where he performed with great distinction.*

I will talk from the perspective of the AHRQ, which I suspect is slightly different than what you will hear from some of the other Federal Agencies. The trials that we fund tend to be in the area of quality improvement, which we understand is not for discussion at this conference, but when trying to improve quality, we are left with the problem of having to start with a baseline definition of usual care. An increasing role of AHRQ is to try to synthesize evidence from the existing research in order to answer questions posed to us by our stakeholders (e.g., health plans, policymakers, and research funders). The questions they want answered are often different from the questions that trials are designed to answer.

This framework comes from Brian Haynes (today's lunch speaker) and originally from Archie Cochran. To answer the progress of information to practice, three questions must be answered:

- Can something work?
- Will something work?
- Is it worth it?

Often our stakeholders are asking the last two questions – whether it will work in their setting and whether it will be worth it. There is often a tension in clinical trial design between the desires that give us a good answer to the "can it work" question – efficacy trials in which the emphasis is on controlling random and nonrandom variation and controlling bias, and therefore often mitigate against using usual care and against the "noise" that Dr. Wennberg talked about – versus the trials that are aimed at answering those questions further downstream – whether something will work and whether it will be worth it. Further downstream, it is important to frame something in the context of what is realistic in practice and how things work in the real world.

In thinking about answering the question "will something work," we often turn to the types of trials that Dr. Wennberg alluded to – whether one calls them effectiveness trials or large practical trials – where one tries to get a better handle on the real world by recruiting larger numbers and a greater diversity of practices and patients. Large practical trials can include placebo interventions but there will always be some element of usual care because, in the attempt to expand the patient population and expand the generalizability, one tries not to tightly control all the interventions in the control arm. An

attempt to overcome the problem of noise and variation in those trials can be made by increasing the sample size of the population.

Our stakeholders often come to us with the question of "is something worth it?" That can be framed as an economic question – is something worth it based on some measure of cost effectiveness or return on investment – but also by asking whether the benefits of making this change will exceed the downsides (potential costs or complications) of those trials. One must frame this question against some comparison group, which must be tied back to the real world of practice. But it is often a challenge to ask whether that appropriate comparison is the current state of practice with all the variation Dr. Wennberg referred to or whether the appropriate comparison is the optimal state of practice, which we realize may not exist frequently. For example, when asking the question of whether bariatric surgery is effective and whether it is worth the benefits of weight loss versus the risks of surgery, it is important to consider whether that is compared to the typical provision of behavioral weight loss programs that patients may get or whether it is compared to the optimal set of intensive weight loss interventions that may not be widely available but may actually be more effective and cost-effective for morbid obesity.

Two cautions or challenges to the use of usual care that AHRQ has encountered:

- Important variations have been demonstrated in care that does make a difference – for example, proven effective care variations are often under-utilized. In comparing something to usual care, we need to be aware of the fact that usual care may be suboptimal and clearly deficient. We should not assume that just because an intervention is better than usual care that it is the best way to improve practice. One example is some work AHRQ did with CMS, looking at new interventions for improving hypertension care. A new technology was proposed to provide specific advice on adjusting hypertension medication. When they took patients in the community who had resistant hypertension, they found that using this device to guide treatment led to important reductions in blood pressure. However, the comparison group was referred to specialists and they showed almost comparable reductions in blood pressure. The point is that the best way to improve hypertension outcomes may not be through a new technology; it may be through improving our patterns of care as they already exist without that technology.

- The second challenge when using a usual care group is that it makes it difficult to synthesize information across studies, which is one of the activities AHRQ supports. Pooling the results of trials, even when they are well-done randomized trials, is difficult when the baseline comparison group is changing among studies. At a minimum, we need to be careful to report and document what goes into that usual care and the attendant variation. This is especially important given the interest across all the Federal Agencies to look at whether the variation in usual care follows patterns of health disparities. In looking at a large trial, one might be tempted to include that "the intervention works better in blacks than in white

patients," but it may be that the pattern of usual care that those patients get differs in some systematic way.

Usual care can be very important in answering questions of effectiveness and of value for real-world decisions, but we need to be very careful to clearly understand what we mean by "usual care" and to document that variation in usual care, both across settings and across specific patient groups.

Louis Fiore, M.D., M.P.H., Department of Veterans Affairs

Dr. Fiore is director of the Department of Veterans Affairs Cooperative Studies Program Coordinating Center in Boston. This research unit is charged with developing clinical trials that will test questions from the field in the context of large-scale, multisite clinical trials carried out in hospitals supported by the U.S. Department of Veterans Affairs (VA).

I would like to describe the research program from the VA, and then comment on the VA's issues with usual care, which have been coming up more frequently in recent years because the concentration of efforts has changed from looking at specific interventions to delivery of care. It is in delivery of care where usual care becomes a problem.

The VA consists of one very large division, which takes care of veterans' healthcare needs, and one very small division, which conducts research. Research by that small division is relevant to veterans and serves two purposes – to supply information that is helpful in treating veterans and to attract and retain physicians to care for those veterans at the VA medical centers. The research done by the VA is organized by the Office of Research and Development and is in four branches – rehabilitation medicine (of crucial interest to wounded war veterans), health services research (helpful to administrators to design how to deliver care), bench work (similar to R01 awards), and the clinical realm.

The Cooperative Studies program is the primary component of the clinical realm. This program operates with five coordinating centers that take an idea or concept from the field and develop that concept into a large-scale clinical trial. The VA has 175 hospitals, 120 of which have Federalwide Assurance (FWA) approval to conduct research. When an investigator in the field has a concept, it is brought to a coordinating center (such as one that I direct), then a full protocol is developed and presented to a peer review committee, and that protocol is then approved and funded or not. Of the 5 coordinating centers, there are 20 or 30 ongoing studies at any one time, involving 10 to 15 hospitals (small studies) to 70 to 80 hospitals (large studies). Currently there are many studies in planning or being closed out in which data analysis is being conducted.

We try to do studies that otherwise would not be done. For example, in rheumatoid arthritis, some very expensive drugs have been approved that inhibit tumor necrosis factor (TNF) and these drugs have widely replaced the less-expensive drugs that are perhaps a combination of two or three pills – replaced because physicians chose the

newer drugs over the conventional drugs. However, the more expensive new drugs were never compared to the less expensive drugs; they were compared only with placebo or a single agent. What someone really needs to do is take the expensive drug and compare it in a legitimate way with triple-therapy drugs that are a fraction of that cost. No drug companies will sponsor that study; those are the kind of studies we try to do.

We tried to design a diabetes study a few years ago, recognizing that one-third of diabetics are depressed. Rather than treating their diabetes any better, we tried to give them some mental health care. After 2 days we decided not to conduct the study because we could not recognize the issues; we did not realize that we did not know enough about usual, enhanced, best possible care for diabetes upon which to add the mental health care. We were totally off the mark and did not know why.

This past month we received a very good score on this study: in patients with COPD, adding a therapist who calls the patient to make sure they have a home bronchodilator, antibiotics, and corticosteroids, and keeps that patient plugged into the system, compared to best minimal care (others might call it proven effective care). We learned a lot in the past few years and we can actually think about the issues. That is the value of a conference like this.

Chad M. Heilig, Ph.D., Centers for Disease Control and Prevention

Dr. Heilig manages the Human Research Protection Office at the U.S. Centers for Disease Control and Prevention (CDC), within the Office of Scientific Regulatory Services, Office of the Chief Science Officer.

I will describe the CDC's interests in medical clinical trials. The CDC sponsors a relatively small number of clinical trials to support preparedness and health promotion efforts. The CDC currently has about 150 projects registered at clinicaltrials.gov, many of which do not entail medical interventions.

The role of direct medical intervention is relevant to several dozen CDC-sponsored projects. Those several dozen trials comprise four categories:

- Prevention efforts and screening (chemoprophylaxis against HIV transmission, development of diabetes, or use of topical microbicides)
- Disease eradication and overcoming drug resistance (two key examples being tuberculosis treatment and malaria)
- Safety (e.g., of immunization regimens and topical microbicides)
- Immunogenicity (e.g., anthrax vaccines and pneumococcus vaccines)

The prevention and eradication trials tend toward pragmatic ends and are frequently conducted in international settings, where there is a higher burden of infectious disease morbidity and mortality and where the concept of usual care is radically different. The

safety and immunogenicity trials tend toward explanatory ends, so as to better understand the mechanics of prevention and eradication interventions.

CDC-sponsored medical trials tend not to include complex medical interventions, in part because the mission of public health requires methods that may be widely and efficiently deployed. The ethical conduct of CDC-sponsored medical trials hinges on framing the research question squarely in the pragmatic, including the appropriate choice of comparison groups. The CDC has famously faced this issue in a well-known placebo-controlled trial for preventing mother-to-child HIV transmission. For the same reason, the CDC is therefore pleased to participate in the discussion of usual medical care as a comparator in clinical trials.

Amy P. Patterson, M.D., NIH

Dr. Patterson is Director of the NIH Program on Clinical Research Policy Analysis and Coordination (CRpac) and Director of the Office of Biotechnology Activities (OBA), both within the Office of Science Policy, Office of the Director, National Institutes of Health (NIH). In the first capacity, she leads a program which provides a focal point for streamlining, coordinating, and harmonizing Federal policies concerning the conduct and oversight of clinical research. In directing the OBA, Dr. Patterson oversees the management of several programs concerned with science, safety, and ethics in a number of critical fields of biomedical research. In addition to her management responsibilities as Director of OBA and the CRpac Program, she is on the clinical staff at the NIH Clinical Center and maintains an active basic research program at the National Heart, Lung, and Blood Institute.

NIH's interest in this conference is that we not only fund trials, we often conduct them ourselves. We have a deep interest and obligation in so doing to advance the science and to see the science applied to the improvement of the day-to-day health and wellbeing of all people. We are here today not only as cosponsors of the conference but to listen and to learn. We hope to emerge from this conference with a set of overarching principles and a conceptual framework about how to intelligently approach trial design and knowing better when, how, and why to factor in usual care considerations.

Marcel E. Salive, M.D., M.P.H., Centers for Medicare and Medicaid Services (CMS)

Dr. Salive is director of the Division of Medical and Surgical Services within the Coverage and Analysis Group of Centers for Medicare and Medicaid Services in the U.S. Department of Health and Human Services. This division is responsible for developing and maintaining coverage decisions for all Medicare beneficiaries using a rigorous and open evidence-based process.

Under the leadership of Mark McClellan, CMS is moving past merely being a regulator and purchaser of health care to recognizing ourselves as a public health agency.

Some of the factors we take into consideration at CMS when deciding whether to pay for new technology, procedures, and services are effectiveness (whether the technology

improves health outcomes) and whether it is better than what already exists. CMS is particularly interested in the latter consideration. We can cover off-label indications for treatments, so there is some room for us to look at evidence in a different way from some of our sister Agencies.

As far as the broad-brush strokes of our coverage policy, for CMS to pay for something there are five things that need to happen, two of which are outside of our scope. Congress determines the benefit categories of what we can pay for. The best example currently is the new Part D drug benefits. This year, we are not paying for those drugs but, as of January 1, 2006, we will be paying for them. That is an example of the benefit category. We require FDA approval for those things that are FDA regulated. We look at coverage issues using a rigorous evidence-based process. And the last two steps are coding and payment.

While the evidence-based coverage process is straightforward and similar to what we have been hearing and discussing about evaluating clinical trials and clinical evidence, the payment side has seen some innovations recently to encourage comparisons to usual care. For example, add-on payments can be made for a service that is demonstrated to be a substantial clinical improvement over what is already available, which encourages developers to perform comparison trials against the current standard of care.

In terms of evaluating evidence in the coverage process, a number of other issues are relevant to this conference. We look for evidence of improved health outcomes. We also look for whether that evidence is generalizable to the Medicare population and whether it can it be applied in general practice to the clinical care of our population. Some recent developments in the coverage process are helpful in this endeavor – coverage with evidence development is being used selectively in Medicare to cover promising innovation where there is insufficient evidence to grant full coverage for the Medicare population. This coverage has been used selectively in areas such as implantable defibrillators and off-label cancer drugs, tying limited coverage to prompt data collection in order to give access to patients as well as collect evidence on the effectiveness of those treatments.

Bernard A. Schwetz, D.V.M., Ph.D., Office for Human Research Protections (OHRP)

> *Dr. Schwetz has been the director of the Office for Human Research Protections in the U.S. Department of Health and Human Services since 2004. Before that time, he was acting director of the same office. This office monitors programs at more than 10,000 DHHS-funded universities, hospitals, and other medical and behavioral research institutions in the United States and abroad. Before going to OHRP, Dr. Schwetz was senior advisor for science the U.S. Food and Drug Administration and there chaired the FDA's institutional review board.*

From the standpoint of OHRP, this is a very important conference to us and to all members of the research enterprise. Any issues that bring together bioethics and good experimental design in science with the questions of compliance with regulations are

both difficult and controversial. I am hoping that this conference will serve as a model for how controversial issues can be brought to discussion. OHRP is a regulatory agency, and as such wants to see that the community doing the research understands the regulations and their relationship to what can and cannot be approved and what relates to being or not being in compliance.

Here is a summary of some pieces of the regulation related to the responsibilities of investigators and the responsibilities of IRBs. Our regulations state, among other things, that the IRB shall be sufficiently qualified through experience, expertise, and diversity of its members to promote respect for its advice and counsel in safeguarding the rights and welfare of human subjects. In addition to possessing professional competence necessary to review specific research activities, the IRB shall be able to ascertain the acceptability of proposed research in terms of institutional commitments and regulations, applicable law, and standards of professional conduct and practice. In accordance with these regulatory requirements, an IRB should have members who can assess the scientific design of the research being proposed and the acceptability of the proposed research interventions in comparison to current routine clinical practice.

We have heard this morning that “routine clinical practice” is sometimes difficult to define, but that is not always the case and it cannot be dropped as an issue just because it is difficult.

Furthermore, in accordance with our regulations, when an IRB lacks the necessary expertise relevant for the review of a particular research project, the IRB may in its discretion invite individuals with competence in special areas to assist in the review of these issues. We need to keep reminding IRBs that that can be done and that they should do it when it would be helpful. These consulting individuals cannot vote but their participation in the meetings needs to be recorded in the minutes of the IRB meeting.

In order to approve research covered by regulations, DHHS regulations require that two things must happen. The risks to subjects must be minimized by using procedures that are consistent with sound research design and that do not unnecessarily expose subjects to risk. Secondly, risks to subjects must be reasonable in relation to anticipated benefits, if any, to the subjects and the importance of the knowledge that may reasonably be expected to result. In order for the IRB to make determinations under our regulations, the IRB must receive and thoroughly evaluate sufficient information describing the research design. Such information must include information adequate to assess the risks and potential benefits of each of the interventions, for each arm of a clinical trial, relative to concurrent routine clinical practice outside of the research context.

Ensuring that sufficient information is received and reviewed by the IRB is a shared responsibility of the investigators proposing the research and the reviewing IRB. If investigators do not present the information to the IRBs that allow them to look at the experimental design in the context of risk and benefit and the IRB does not ask for that information, that particular issue may not be reviewed in sufficient depth. If the investigator presents the information to the IRB, I assure you they will look at it; that is

the nature of IRBs. If the investigators do not present the information about the research and its experimental design in relationship to subject safety and the IRB does not ask for it, then the protocol and the experimental design will likely be under-reviewed.

In multisite studies, how can many IRBs review a protocol and miss something? If it is "something" that is a regular topic of discussion, multiple reviews do pick up those issues, even if missed by one IRB. Issues that are controversial or are not particularly discussed at PRIM&R and other meetings where IRBs get together or investigators get together to talk about issues of design in relationship to risk, those issues can be missed by many IRBs even if they are all looking at the same protocol. It is not a given that everything will be picked up if enough IRBs see it, at least from our experience.

This is an important shared responsibility between IRBs and investigators to ensure that the compliance issues are well understood and followed, so that we adequately protect research participants.

Robert J. Meyer, M.D., Food and Drug Administration (FDA)

> *Dr. Meyer is director of the Office of Drug Evaluation II, in the Office of New Drugs, Center for Drug Evaluation and Research, at the FDA. This office comprises the Division of Pulmonary and Allergy Products, the Division of Anesthesia, Analgesia, and Rheumatological Products, and the Division of Metabolic and Endocrine Products. He came to the FDA from the Argonne Health and Science University in Portland, where he was assistant professor of medicine in the Pulmonary and Critical Care Division.*

The gold standard for drug trials is the RCT, with an emphasis on testing against a control. The intent is to minimize systematic variability. The FDA does not have the legal authority to require comparative trials against existing therapy; although comparative trials are often conducted. In choosing the comparator, it is important for the FDA that the trial results end up being interpretable. In Phase 3, if a trial is not sufficiently well designed, the FDA can put it on clinical hold, even if it is not overtly unethical, since a trial that is not of sufficient design to answer the question posed puts people at risk without the benefit of participating in potentially meaningful research.

The FDA is an evidence-based agency. We do not wish to and cannot control the practice of medicine. We do want to provide important information to inform the practice of medicine, but it is neither the FDA's intent nor its purview to control the practice of medicine.

Discussion

Q: For the OHRP: Do you have any feelings about central IRBs for multicenter trials?

Dr. Schwetz: We support whatever form of IRB review is appropriate for the research study. That might be a central IRB or it might be a local IRB. We have made it clear that there are ways that central IRBs can be used and be fully compliant with our

regulations. We support the concept, and will be having a meeting on Thursday and Friday of this week to explore further how the models that are available for research review can be best matched to the kind of review – social and behavioral or biomedical, whether single site or multisite. We support whatever kind of IRB is appropriate for the research; in some cases a central IRB is appropriate.

Design Considerations in Randomized Controlled Trials With a Focus on Usual Care Arms: Compared to What? Some Thoughts on the Choice of Controls

Janet Wittes, Ph.D.
Statistics Collaborative, Inc.

Design Considerations in Randomized Controlled Trials With a Focus on Usual Care Arms: Compared to What? Some Thoughts on the Choice of Controls

Janet Wittes, Ph.D., Statistics Collaborative, Inc.

Dr. Wittes is president of Statistics Collaborative, Inc. She is a member of many advisory committees, including a large number of data and safety monitoring boards (DSMBs) for randomized clinical trials sponsored by industry and the government. Her own research focuses on the design and analysis of RCTs. From 1990 through 1995, Dr. Wittes served as editor-in-chief of the journal Controlled Clinical Trials.

In this talk, I will discuss a taxonomy of control groups, because of its direct relevance to the problem of usual care arms. As Thomas Pynchon said in "Gravity's Rainbow" – "If they can get you asking the wrong questions, they do not have to worry about answers." I will argue that the choice of the control group in a trial defines the question the trial is asking.

Imagine a new or experimental therapy or a new use for an old therapy. In trying to learn about its effect, we might be tempted to ask the "wrong" question: "Does it work?" The real question we should address instead is: "Does it work compared to x?" where x is a specific control group. The question posed at this conference is the definition of x in that question of "compared to x"; the operative problem is how to choose the comparator. And, of course, the x we are most interested in here is "usual care."

If patients volunteer to participate in a clinical trial, the treatment received by those in the control group should generally be at least as effective as they would receive under ordinary care. My statement may sound precise, but really it is quite vague because we may not know what "ordinary care" is, we may not know what "effective" is, and we may not know what "at least" is. But in some sense, we all believe that patients should not be penalized for being in a control group in a clinical trial. And not only should patients not be penalized for being in the trial, their participation should be expected to lead to increased knowledge for the medical community as a whole. Therefore, the trial should be asking a meaningful question; at the end of the trial, the reader of the study report should be able to say, "I know what question the trial asked." Having a meaningful question is necessary, but not sufficient, for a trial. We should require that the answer be interpretable. By "interpretable" I do not necessarily mean "unequivocal"; we are at the mercy of the data and sometimes answers from a trial are not clear. But we should be able to interpret the answer – the study should tell us whether the intervention worked compared to x, whether it didn't work compared to x, or what more information we still need.

This conference deals with control groups defined as "usual care." First, we must come to some consensus about what we mean by usual care. I will start with some general scenarios that define usual care. The easiest case, which we statisticians love, is the situation in which no treatment is available. Almost as easy is the case of reversible symptoms – people who have symptoms are willing to endure them for a short time to see whether a new therapy, compared to none at all, is effective. In some situations a standard of care, promulgated by some process of consensus building, is available –

one example is a lipid-lowering treatment for people with high levels of LDL-cholesterol. Such treatment is a (nearly) universally accepted approach to therapy or prevention. In the more typical situation several different standard treatments are in use; they may differ by hospital, differ by region within "wealthy" countries, or differ between wealthy and poor countries. "Usual" care, in contrast to "standard of" care, is what is "usually" done. "Usual care" is often situational – it may or may not reflect a general standard of care.

To define appropriate control groups, I will mention each of the above situations in turn starting with the "no treatment available" case. For example, one can easily design a trial for a new vaccine where none is available. I work with malaria vaccines. In our trials, we randomize some people to an experimental malaria vaccine and others to a control group that receives rabies vaccine or some other vaccine that has nothing to do with malaria. In these trials, everyone receives a vaccine. At the end of the trial, we give the experimental group the control vaccine, so they receive some benefit from participating, and we promise participants in the control group that if the experimental vaccine proves safe and reduces the incidence of malaria, they will receive the malaria vaccine. In such a trial, the question is transparent: "Does the new vaccine prevent malaria?" The analysis is conceptually simple: counting the cases of malaria in the two groups answers the question. (Of course, the actual counting of malaria cases is not simple, but that is another story.)

Replacement therapies for genetic diseases (for example, alpha anti-trypsin for PiZZ emphysema) are clear – participants are treated with a replacement enzyme and no other replacement is available. Similarly, studies of prevention of hard clinical outcomes – statins for MI, estrogens to prevent coronary disease in women (which did not work), beta-carotene to prevent cancer (which also did not work) – clearly ask whether an intervention reduces the incidence of a specific endpoint compared to placebo.

For situations in which an effective therapy or preventive agent is available, we need to compare the new therapy to it. The Declaration of Helsinki, Paragraph 29 states, "[T]he benefits, risks, burdens, and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods." But it does not tell us what it means by "best." It goes on to say, "This does not exclude the use of placebo or no treatment, in studies where no proven prophylactic, diagnostic, or therapeutic method exists." This sentence suggests that the Declaration *does* exclude the use of placebo or no treatment in studies where no proven prophylactic, diagnostic, or therapeutic method exists. I bring this up because this statement has led to a lot of confusion. The fact that a trial has a placebo group does not mean the control group is receiving suboptimal care. A "usual care" study often compares "usual care plus placebo" to "usual care plus experimental therapy." When we think of "placebo" we must distinguish among different kinds of placebo-controlled trials.

In the situation with no available treatment, the logical control is either no treatment or a pure placebo – a placebo not against the background of something else. In those cases, a legitimate question is why even bother with a placebo; why not simply have a no-treatment control? The reasons have to do with the technical feasibility of

implementing a trial and making unbiased assessments of outcome: a placebo may be necessary to encourage compliance (if you know you are in a study but not being treated, why bother to come back for your visits?) and to reduce bias in assessing endpoints. Placebos are generally useful, but not always necessary. In fact, in some cases it may be useful to have both a placebo-control and a no-treatment control. For example, studies of oral mucositis typically have three groups. In the active and placebo arms they swish a solution in their mouths. Concern that the vehicle constituting the placebo solution may be somewhat harmful or somewhat beneficial has led to the use of a true no-treatment control.

The Helsinki Declaration goes on to say: "Placebo may be ethically acceptable if a proven therapy is available for compelling and scientific reasons, or where a prophylactic, diagnostic, or therapeutic method is being investigated for a minor condition and the patients who receive placebo will not be subject to any additional risk of serious or irreversible harm." The key words there are "serious" and "irreversible." In studies of pain, for example, pure placebo may be used because pain, at least at moderate levels, is neither serious nor irreversible. A trial may randomize to a new analgesic or placebo for joint pain or migraine. The protocol specifies a rescue therapy that the participant may receive if the level of pain becomes unacceptable.

At the other extreme, placebos are used in some trials of severe psychiatric illness – clinical depression, hospitalized mania, and schizophrenia. These trials often are performed within hospitals by removing people from their drug, watching them carefully to prevent them from harming themselves, and testing against placebo.

Trials that use pure placebo or no treatment ask implicitly or explicitly whether the new treatment works better than placebo or no treatment. But what if a standard of care is available that has been shown in randomized trials to be effective, for example, antihypertensives to prevent stroke or protease inhibitors for HIV/AIDS. No one would perform a long randomized trial taking antihypertensives away from patients with uncontrolled high blood pressure. In those cases the trial compares the standard of care plus a placebo to the standard of care, or usual care, plus the experimental therapy. That, too, is a placebo-controlled trial. It addresses the question of whether a new therapy added to the standard of care improves outcome; this type of trial does not investigate replacing the standard of care with a new therapy.

What about the case in which a standard of care is available but it has not been shown to be effective or necessary? We heard early today that antibiotics for otitis media had not been shown effective, so a controlled study against placebo would not put people at known excess risk, even though there was a standard of care. Before the Women's Health Initiative, many of us thought we knew that estrogen use prevented myocardial infarction, but the therapy had not been shown effective and therefore a trial was important.

If the standard of care is not evidence-based so there is some kind of clinical equipoise among the clinicians and the participants in the trial, it is reasonable to think about studies that compare standard of care to a new therapy. If we need to blind the trial, we

may require a placebo control. Here the question is different: the trial is asking whether the new therapy is better than, or different from, the standard of care. If the study shows the new therapy to be beneficial, the medical community can replace the standard of care with the new therapy.

A more difficult question related to standard of care is the following: is the new therapy equivalent to the standard of care (whatever "equivalent" means)? Presumably, if the new therapy has equivalent benefit but has some other advantage – it may be easier to administer, it may be less expensive, it may have fewer side effects – it will become the treatment of choice or at least one of the treatments of choice. I prefer to refer to "equivalent" as "not very different from" and its sister "non-inferiority" as "not unacceptably worse than." In this type of trial, we define the equivalence of a new therapy and the standard of care as being the same within some predefined margin. The problem with this kind of study is that we cannot determine from the data of the trial itself whether the standard of care worked in this particular trial. Bob Temple talks about studies in clinical depression in which the investigators randomize people to a new antidepressant versus one that has already been shown to be effective. On finding no difference, they conclude, "My new drug has fewer side effects and is not worse that the standard. Therefore, we should replace the standard therapy with it." He points out that, in the context of diseases that ebb and flow, where there are regressions to the mean or what some people would call placebo effects, one cannot know that, in a particular context, the therapy that had been proven effective in previous randomized trials was actually effective in this setting. Perhaps the sample size was not large enough to detect differences, the study may have been performed less than optimally, the patient population may have differed meaningfully from the populations previously study, or the background therapy may not have been the same as in previous trials.

In the absence of a standard of care, the protocol can define a regimen it calls "usual care" for the trial. It might provide a menu of protocol-specific options from which the investigator chooses. Later this afternoon we will hear of a trial in which the menu had two items. The more similar the "entrees" in that menu, the less variability in the control group and the easier it will be to compare the new treatment to the permissible mix of options. But of course if the options are small garnishes on a main dish, having a menu does not provide a wide choice to the investigator.

Another option is to leave the choice of care completely to the clinicians, in which case usual care is defined as what the clinician normally does or whatever the particular community does. Such a trial has a much more heterogeneous control group than does either of the above scenarios.

These designs ask somewhat different questions. When the protocol defines the regimen it calls "usual care," the question is how the new treatment compares to that specific regimen. Such a trial, by addressing a homogeneous question, increases its statistical power. But the answer such a trial produces may not be informative because, if that protocol-defined regimen is one that nobody or very few people use, the answer may have little relevance. John Tukey, one of our greatest statisticians, said, "Far better an approximate answer to the right question, which is often vague, than the exact

answer to the wrong question, which can always be made precise." Using a protocol-defined regimen provides us with a precise question. The right question, however, addresses usual medical care – what is done in practice – which is by nature vague, that is, heterogeneous and perhaps imprecisely defined. A clinical trial that compares the experimental therapy to some artificial "average" of therapies will give us approximately the right answer. In order to incorporate into the design the variability in the control group, the sample size will certainly be higher than the sample size of a trial that protocolizes the control.

And please don't hold out the hope that at the end of the trial anyone can tease out whether the treatment works against specific type of controls. First, the sample size will be too small for regimen-specific comparison and, more important, if the study is an unblinded study of usual care against experimental therapy, we will not be able to determine what usual care regimen would have been used for particular patients in the experimental group. Some trials attempt to ask the physician what regimen they intend to use, and only when the regimen is defined do they randomize the potential participant. This method allows unbiased comparison of true "intent-to-treat"; analyses from such trials should be based on the planned, not the actual, treatments.

I want to distinguish what we call "usual care" in clinical trials to usual care that occurs in practice. So-called usual care in a protocolized definition is often a fixed dose in the control group compared to a fixed dose in the experimental group. In practice, often the doses are titrated and people are treated to target. Of course, some clinical trials do treat to target. Hypertension trials do so by randomizing to a target goal hypertension; participants in the treatment arm get their doses adjusted to keep blood pressure at the level of the target while the doses in the control group change in response to fake data.

An example is oral mucositis, mentioned above. Chemotherapy and radiation therapy may cause erythema and ulcers in the mouth and throat. The FDA guidance in these trials recommends three groups: treatment plus vehicle (a solvent), vehicle, or usual care. The participant swishes the solution in the mouth and spits it out. The vehicle is supposed to be inactive. The active agent in the treated group is dissolved in the solvent. The comparison of the treatment plus vehicle to the vehicle itself is blind. The problem is that the vehicle is not inert. Not being water it may have activity in itself. Therefore, the FDA has mandated that such studies have a usual care group as a second control.

The arguments for this control group are two fold. Suppose the vehicle is effective but treatment plus vehicle is even more effective. Then, without a no-treatment control, the study cannot produce a good estimate of the magnitude of the effect of treatment because some of the efficacy was due to the vehicle. The other argument is that the vehicle might be aggravating erythema and then the active treatment might reduce that aggravated erythema. So comparing the treatment plus vehicle to vehicle-only would lead to the conclusion that treatment is better than vehicle. In the absence of no-treatment control, one wouldn't know that the treatment just counteracts the negative effect of vehicle. The problem, of course, is that assessment of degree of erythema and

ulceration is somewhat subjective so that the no-treatment control has the potential for bias.

Tukey says, “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.” Sometimes we just cannot answer what we want to answer!

There is another type of trial that I have not addressed – a trial that looks at dose response for a proven therapy. This type of trial asks whether to change the dose. An example comes from the early trials of AZT – the first trial, which used a rather high dose, showed efficacy. Further trials led to decreasing doses. Similarly, trials may be designed to study biological targets. Trials of transfusion triggers provide another example – how low can the transfusion trigger go? In the discussion of ARDSNet that will occur later in this conference, speakers will address the question of how separate the doses or intensity should be in a trial – the farther apart the interventions are, the more the trial can maximize the power for the clinical endpoints. If the effect of dose is monotone, comparing the lowest effective dose to the highest safe dose will give the most statistical power. The more relevant medical question is often to compare the lowest commonly used dose to the highest commonly used dose. The differences between these two doses will be smaller. As the triggers or doses become closer to each other, the power to detect differences will decrease and the sample sizes will necessarily increase.

A few caveats: I have only brushed against the important question of generalizability. Studies have subgroups of patients within them. People always ask about the relevance of the treatment to various subgroups of patients within the population. Even more difficult is the relevance of the inference we make to subgroups of patients not studied at all. I have alluded to, but not fully discussed, subgroups of treatments within the usual care arm. Finally, I have not even mentioned the problem of choosing the appropriate endpoints.

In summary, the choice of control is crucially important to the implementation of a randomized clinical trial and to the inference from it. Different controls imply different questions. In evaluating the inference from a trial, look at a control group and make sure that the questions being asked are consistent with that control group. To end with a warning from Claude Bernard, “When you do not know what you’re looking for, you will not see what you have found.”

Discussion

Dr. Levine: When Albert Einstein was asked by the U.S. Senate why scientists keep giving them the wrong answers, he replied, “Scientists always give you the right answers, but they are to the wrong questions.”

Q: In the usual care subgroup, you talked about the loss of power within the subgroups, which is an issue. But sometimes the additional problem is that, in the treatment group, you do not know what part of that subgroup fits the control – if the control has a bunch

of different groups, you do not know who in the treated group matches the subgroups in the control arms.

Dr. Wittes: That's absolutely right. I was thinking of the simpler situation in which you could do the match. Typically you cannot.

Q: You have talked about the control group with regard to the question and referred to both explanatory efficacy and pragmatic effectiveness trials. If we are asking a question about how we should alter practice, you have emphasized the importance of having a usual care group. Can you comment on the sequence of trials? For example, there are a number of trials that have been conducted as effectiveness trials prior to the identification of efficacy. When those trials produce no difference between the groups, the community is left with the same question with which it began. Without knowing first that efficacy is present, that an explanatory trial has been successful, one is hard pressed to justify the execution of an effectiveness trial. And yet one sees them conducted by very good groups – for example, a recent trial conducted by the Canadian Clinical Trials Group (CCTG) in pulmonary artery catheter use, which left us at the end with the same sense that we had at the beginning. Would you speak about the sequence?

Dr. Wittes: That is a very difficult question. I first have to make a confession – I have a hard time distinguishing effectiveness or efficacy trials, except at the extremes. If you give me a trial and ask if it is an efficacy trial or an effectiveness trial, I often scratch my head. I want to think, instead, of proof of concept trials – that I have a theory and I have a mechanism and I want to see if that concept works. That is what other people call "efficacy." And then, does it work in the real world? The problem is that there are different kinds of real worlds and the real world is becoming increasing heterogeneous. It depends a lot on the clinical setting. I work with orphan drugs and on malaria and heart disease. For some conditions there are 21 people in the world; for malaria there are millions and for heart disease there are millions. My own view is colored by thinking, "In heart disease if you want to do a trial of 5,000 you do it, and then another of 10,000 and another of 60,000; there is no big deal". There I think the sequence is not only important but doable. In situations where the patient population is much more limited, then you have to skip some of those steps and hope you can interpret at the end – and sometimes you cannot.

Q: The only thing to add to it, if you are a taxonomist, is switching between standard care and standard of care and usual care. In screening trials, usual care can be nothing, standard care is at least something that someone does, and at least in theory standard of care is broadly acceptable care. Those definitions are not standardized. People lump together things that are broadly dissimilar under those taxonomic labels. So there is a fundamental problem of using the labels themselves without proper definitions.

Dr. Wittes: I totally agree and I struggled with that. As you noticed, that was part of my "usual care versus usual care." The language is not clear and not unambiguous, and we all use it in different ways. The way I think about it is, there is something that most

people do that has been shown effective from clinical trials – it still may not be a standard of care because people may not use it. My college roommate was a botanist and taught me to think in terms of classifications and how difficult they are.

Dr. Fiore: The question that a trial is meant to answer is often a different question than what the clinician wants to know. For example, Drug B is expensive and does not require monitoring; Drug A is standard (everyone with this condition gets Drug A) but it is a titrated drug that is difficult to monitor. In a head-on trial, Drug B appears to be superior because, in the trial, no provision was made for super-vigilant monitoring. It is known in the community that monitoring is less than optimal because it is tedious, expensive, and time consuming. I am a public health person and I decide that Drug B is superior to Drug A because, in my community, therapy will be delivered adequately and there will be fewer bad events for the cost of the drug. I am a clinician, and I want to know whether Drug A, if given precisely, is as good as Drug B because my patient does not have the funds to buy Drug B and I do not want to be using Drug B. So there are two very different goals. Could you comment on that, as it speaks to the issue in a lot of these studies?

Dr. Wittes: In your presentation, you said that the VA is able to do back-to-back comparisons, which I think is very important. A lot of drug trials are against placebo for symptoms and are not back-to-back. When I was involved in the RALS trial for heart failure, I observed that you have to monitor potassium carefully. In the trials they would monitor it very carefully and there was a 30 percent reduction in mortality. Out in the community, all of a sudden people started dying of potassium imbalance. My answer is an example of how difficult it is. You do not want to take a trial and say to the doctors, "Don't bother; in this trial we want to see how the treatment will work if you don't monitor particularly well." To me, those are questions that in some ways we cannot answer. You cannot do the experiment to answer the question of how people will actually behave. The only hope is that, if you really know that things need to be monitored, you will educate people. I know that is not a really good answer.

FDA Perspectives on Usual Care Control Groups in Regulatory Decisionmaking

Robert J. Meyer, M.D.,
U.S. Food and Drug Administration

FDA Perspectives on Usual Care Control Groups in Regulatory Decisionmaking

Robert J. Meyer, M.D., FDA

Dr. Meyer is director of the Office of Drug Evaluation II, in the Office of New Drugs, Center for Drug Evaluation and Research, at the FDA. This office comprises the Division of Pulmonary and Allergy Products, the Division of Anesthesia, Analgesia, and Rheumatological Products, and the Division of Metabolic and Endocrine Products. He came to the FDA from the Argonne Health and Science University in Portland, where he was assistant professor of medicine in the Pulmonary and Critical Care Division.

I would like to reiterate the point I made earlier: while it is of great interest to those who are working in the FDA to think about how what we do would be best put into practice, it is neither the legal purview nor our regulatory mission in the strict sense to design trials to inform the practice of medicine as much as to prove that a drug or a therapy is effective.

I will offer some brief comments about general expectations of the data needed for drug approval; much of these comments would also apply to devices and certainly to biologics. Then I will talk about some regulatory considerations for usual care, much of which echo Dr. Wittes' talk but come from a different angle. Then I would like to use a case example to present some of the challenges that can come into play in the regulatory environment with the issue of usual care.

From the FDA perspective, the approval of a new drug therapy requires substantial data from adequate and well-controlled clinical studies. That requirement is in the Food, Drug, and Cosmetics Act, which forms the legal basis for what we do and is the underpinning of everything we do. As I said earlier, I would like to emphasize the "well-controlled" aspect of this. We look for studies where variability among the patients is thought about and controlled to a reasonable degree, and the design and conduct is done in a way that protects against biases. The choice of a control group is critical to the success of a trial, in terms of assuring that the study is well controlled and that the results are interpretable and can lead to regulatory decision-making.

The International Conference on Harmonization (ICH), which is a consortium of regulators and industry from the United States, the European Union, and Japan, has a number of guidance documents on various topics. The document included in your background information is the ICH E10 guidance. "E" refers to "efficacy," which means a clinical guidance. The E10 document talks about control groups, mirroring the same kinds of considerations in the U.S. law and regulations. It allows for a placebo concurrent control, no treatment concurrent control, active concurrent control, dose-response control, external (including historic) control, and multiple controls. None of these precludes usual care.

To make that point, I will focus on one of these controls, primarily because it is often missed when discussing regulatory studies. Dr. Wittes used the term "pure placebo control." Often at the FDA what we consider to be placebo-controlled studies are actually studies in which the placebo or the test drug is an add-on to some background

of therapy. Many such studies do not use pure placebo controls, and the ethical criticisms of those studies are often misplaced, because people do not seem to understand this fairly simple and fundamental point. For instance, it is often the case that a new chemotherapeutic drug, while tested against a placebo, is added to an existing, accepted regimen for that particular cancer. Or in the case of Type II diabetes, a new hypoglycemic agent is tested against placebo in the background of existing therapy such as sulfanurias or PPARS. In both examples, those receiving "placebo" are receiving standard care.

What are the important considerations in using "usual care" as a comparator? Now I am switching to the situation in which a drug is being tested against a different regimen that meets someone's definition of "usual care," although I am not entirely sure that "usual care" is defined adequately (at least for regulatory purposes). For the purposes of regulatory studies, the question is whether we are talking about something that is "ad lib" (up to local standards and local practice). As someone who did most of my postgraduate training on the East Coast and then went into faculty practice on the West Coast, I was astounded by the differences in the practice of medicine that people believed were well substantiated by data but that were entirely divergent. So there is the ad-lib usual care or a "defined consensus usual care." In the latter, you may get a group of experts together from multiple sites in the United States and come to some consensus as to what might be considered a standard of care or usual care for use throughout the various study sites. The ad-lib usual care, where the practice at each site or by each investigator may vary, renders the interpretability of the results very difficult, and it can undermine the likelihood of the study showing a difference. Yet, finding a consensus may be difficult to achieve.

When usual care serves as the active concurrent control, there are a number of other considerations. If the trial design is a superiority design, the main issue may be proper blinding of the assessment, which could be handled by some sort of double dummy design (but that could become quite complex). Blinding would certainly be an issue and finding a consensus of some reasonable boundaries of usual care could also be a challenge.

If the design is non-inferiority, there are additional issues. In addition to blinding, if you are using a non-inferiority design against usual care, you need to have an accurate estimation of the likely treatment effect from usual care in that experimental setting to allow for proper design and analysis, particularly in choosing the non-inferiority margin and the appropriate number of patients for that study. You also need to be convinced that usual care would reliably show a treatment effect in such a study; the case of the antidepressants was raised in the last talk, and that is an apt one. If you do not know that the study would reliably show a reproducible treatment effect and if you fail to find a difference between your test and your usual care, you will not know whether that means that no treatment difference really existed, or whether that study simply "failed" and neither treatment would have beaten placebo if placebo had been used. This is what Bob Temple and others within the FDA refer to as "assay sensitivity" – without a placebo group, you may not know that the test was sensitive to detect a difference should one exist.

The issue for the FDA is the need for a robust, reproducible estimate of the effect size for usual care, but this frequently is not the case, especially where “usual care” involves multifaceted treatments developed in the practice setting. These practices may be informed by individual pieces of randomized controlled trial data and other data showing that the regimens work separately, but put together, it may be a question as to whether you have a reliable and robust estimate of the effect size if the combination of therapies has never been adequately studied. Without such an estimate, the use of a usual care comparator group as the sole comparison may be difficult unless you want to study the new treatment to establish superiority. In the case of studying the new treatment to establish superiority to usual care with no placebo group, you only have to make the assumption that usual care would not have been worse than nothing. While that may be possible in some circumstances, it is not always the case; an example is the Women’s Health Initiative situation for estrogens and heart disease, where going into the study, most investigators anticipated an estrogen benefit to heart disease, not worsening.

Most modern drug trials seen by the FDA are multi-centered, if not multinational; some are also conducted in developing countries such as in Latin American and the former Eastern Bloc. “Usual care” in one community may not be usual in another. Therefore the selection of a single “usual care” regimen may not be acceptable in some communities. One example is doing a large outcome study of a Cox-2 inhibitor for the treatment of arthritis primarily as a safety trial. If you were doing it in the United States you would want to know how it compares to ibuprofen, since that is one of the most widely used NSAIDS in medical practice and in the over-the-counter setting. However, ibuprofen is not at all common in the European Union, so they might want a different comparator. How do you solve that, in trying to design one trial to address important questions about whether Cox-2s are safer from a cardiovascular standpoint compared to common practice with NSAIDS in the United States and abroad? One possibility is to study more than one “standard of care”, but selection of multiple usual care regimens may complicate the interpretation of the trial.

In certain circumstances, usual care is considered so inferior by some practitioners, some trialists, and some ethicists, that there may be concerns about including it in a study. The following case was not within my drug office, so I know the case reasonably well but not all the details; it is a good example of some of the complexities that usual care can raise.

Eptifibatide is a peptide platelet inhibitor that binds to the platelet’s IIb/IIIa receptor and therefore acts by inhibiting platelet aggregation. It was approved for use in acute coronary syndrome and in the setting of percutaneous angioplasty. After its approval, the sponsor wished to study the drug at a higher dose, because they had developed some pharmacokinetic and pharmacodynamic evidence that suggested that a higher dose of the drug might lead to better effects in terms of preventing coronary occlusion, because of higher and more consistent occupancy of the IIb/IIIa site. The sponsor proposed a placebo-comparator trial in the percutaneous coronary transluminal angioplasty (PCTA) setting. Instead of a dose-comparator trial, they wanted to do this as a placebo-controlled trial that would allow for a more practical study size. They

proposed using eptifibatide before the angioplasty and the stent placement. They proposed an early bailout use of an approved IIb/IIIa agent if there were any suggestive findings after the angioplasty, and would track that use as an efficacy outcome.

This study was proposed not only after the eptifibatide was approved for use in PCTA but also after other IIb/IIIa inhibitors were approved for that setting as well, with important morbidity and mortality findings. These drugs were approved on a combined endpoint of mortality MI or cardiovascular revascularization. Originally, the FDA thought the study proposed by the sponsor was ethically unacceptable – it was a pure placebo study and there was no alternative drug given, although aspirin and some other platelet-acting agents were allowed as background.

After being placed on clinical hold, the sponsor argued successfully that, in the setting proposed (specifically angioplasty with a stent placement), there was only one definitive study in the literature on the use of eptifibatide in that setting and many authorities in the cardiovascular field thought that this study bore repeating prior to it being accepted as the basis for medical practice. The sponsor also provided compelling evidence that usual care in such patients did not universally involve the use of IIb/IIIa inhibitors, due to costs of these agents, the major adverse effects of the agents (bleeding), and the uncertainty of efficacy in certain patients (primarily the low-risk patients, since the data were developed primarily with high-risk patients). Therefore, after extensive discussion with the sponsor, the FDA did allow the study to proceed with “usual care” plus the test drug versus usual care plus placebo in low-risk patients. We insisted that the informed consent make clear to the participants that the IIb/IIIa inhibitors were available and used in some circumstances, so that if they opted to go that route they could do so without going into the study. Parenthetically, we also strongly encouraged the sponsor to publish a description of some of these issues, and I think they subsequently did so, although it took many years. The study was positive and led to new labeling and dosing recommendations, leading to a more effective regimen for the drug.

In summary, usual care can be incorporated into a number of study designs that satisfy the regulatory need. However, the use of usual care groups can present challenges, particularly when used as a comparator as an alternative rather than an add-on therapy. The question remains about the definition of “usual care” in any given circumstance and how universal that is, particularly in the setting of multicenter trials. Is there reasonable consensus and agreement that can be reached, so a relatively homogeneous comparator group can be used? How reliable and defined is the efficacy of usual care – how robust is the evidence and how much do you know about the likely effect size if you were going to enter it into a trial? Commonly, the issue is that a lot of the randomized controlled data underlying usual care may be several years old, and the practice of medicine changes over time. What does that do to the underlying assumptions about the likely effect size? Finally, comparing two disparate groups in clinical trials for drug approval raises the issue of blinding.

Discussion

Q: You described how, using a particular example, the FDA considered notions of usual care in approving a particular trial involving a platelet inhibitor. For the sake of completeness, you need to be aware that the article that was written about those trial design discussions and that was referenced in the handout has also been subject to critical appraisal. I immediately admit a bias because I am a coauthor of an article that was critical of that trial. I encourage members of the audience to read the article that directly addresses the article that you describe, to make a judgment about whether the investigators and the FDA judgments about the trial were defensible. One of the critical things about that article is that, when it came to defining what was usual care, there were considerations of efficacy and harms but also cost. If you look at the survey conducted by the investigators to support the trial after the FDA put the initial clinical hold on the trial, the issue of costs rather than efficacy and harms loomed large. Our contention is that the conduct of the placebo-controlled trial was not defensible, based on the information provided by the investigators to the FDA, including the survey that was conducted.

Dr. Meyer: The other part of the package presented to the FDA was a lengthy discussion on the matter by a prominent ethicist. While that was not the only thing we considered, it was an important part of our consideration.

Q: In introducing that case study, you cast it as a dose-response question. Was it supposed to be cast as a dose-response question? If so, did you answer it or do you want to recast it?

Dr. Meyer: It was not my intention to cast it as a dose-response question. The dose-response issue came up in the case, but the study was not so designed. It was designed to say that the new dose was safe and effective in its own right but not to say definitively whether it was safer or better than the existing dose.

Dr. Haynes: The issue of blinding has come up several times in usual care trials. But blinding is not one particular procedure; there are several levels of blinding. There are manageable levels of blinding in these trials. With objective endpoints like life and death, blinding is not an issue. However, even with subjective endpoints, there are ways to blind outcome assessors, statisticians, manuscript writers, etc. that can at least minimize the difference. I am not sure that should be held out as an absolute reason for not doing usual care trials. In fact, it is probably a manageable issue in usual care trials, and it is not unique to usual care trials.

Dr. Wittes: I did not mean it as an absolute contraindication. I totally agree that the most important thing is blinded, unbiased assessment of outcome. I brought it up various times because it is part of the consideration of the design, but it should not dominate.

Ethical Considerations in Randomized Clinical Trials: Focus on Usual Medical Care and Ethics of Trial Design

Charles Weijer, M.D., Ph.D., FRCPC
University of Western Ontario

Ethical Considerations in Randomized Clinical Trials: Focus on Usual Medical Care and Ethics of Trial Design

Charles Weijer, M.D., Ph.D., FRCPC, University of Western Ontario

Dr. Weijer is Canada Research Chair and associate professor of philosophy, medicine, and epidemiology in biostatistics at the University of Western Ontario in London, Ontario. His interests include research ethics and philosophy of science. In 2005, he assumed the Tier 1 Canada Research Chair in Bioethics at the University of Western Ontario. He has published extensively on the ethical analysis of research benefits and harms, protection of communities in research, and standards in international research. He has been an important contributor to the development of national and international policy in research ethics.

I will share with you some of my thoughts about ethical considerations in randomized controlled trials in general, with an emphasis on the usual care question.

Starting with a broad picture view, it is instructive to note that so many of the issues that we consider the top controversies in the ethics of research today turn on questions of acceptable benefits and harms. Whether we are talking about placebo controls, developing country research, emergency research, or the ARDSNet case, all are about what constitutes acceptable benefits and harms in research. It is instructive to note that, for things like informed consent, there are well-articulated norms and procedures – we have a theory of autonomy and precise procedural specifications for how to obtain informed consent. The problem is that the norms and procedures for the ethical analysis of benefits and harms in research are not widely known. That gives us a causal story for why these issues are so contentious – because disagreement and polarization are predictable outcomes in a domain ruled predominantly by the vagaries of intuition.

We do not only have the vagaries of intuition, we have Federal regulations, which purport to tell us what to do. The Federal regulations on harm-benefit analysis instruct IRBs that risks to subjects must be minimized and reasonable in relation to anticipated benefits of any subjects and the importance of the knowledge that may reasonably be expected to result. How does one do that, faced with particularly complex protocols? There is no widely understood guidance on this question. These regulations are not self-interpreting. Without a conceptual framework to guide interpretation, we will get intuition, *ad hoc* responses, and disagreements about whether benefits exceed harms in any particular case. Therefore, it is important that researchers, IRBs, and OHRP require clear and unambiguous guidance on the ethical analysis of study benefits and harms.

To make the point more definitive that these regulations are not self-interpreting, here are questions that the regulations do not answer. To apply them systematically, we need answers to them. The regulations do not tell us:

- Which risks to subjects must be minimized?
- To what extent must they be minimized?

- Which risks and which potential benefits are to be considered in the reasonableness determinations?
- By what measure does one determine that risks are reasonable in relation to benefits to subjects?
- By what measure does one determine that risks are reasonable in relation to the knowledge that may result?

In the absence of that conceptual framework and that meaningful clarification, no wonder there is such disagreement focusing on questions of benefit and harm in clinical research.

I would like to present a systematic approach that provides just such answers to researchers, IRBs, and the OHRP. It is a refinement of the risk framework developed by the National Commission's report on IRBs, "The Belmont Report," and Professor Levine's work for the National Commission. As these documents served as the basis for the Common Rule, what I will call "component analysis" serves as a basis for the interpretation of the Common Rule. It was endorsed formally by the National Bioethics Advisory Commission in its final report in 2001 and, to my knowledge, it is the only systematic and comprehensive approach to the ethical analysis of research benefits and harms. A paper describing component analysis in detail is in your package (Weijer C, Miller PB. When are research risks reasonable in relation to anticipated benefits? *Nature Medicine* 2004; 10:570-573).

Component analysis starts from the understanding that clinical research often contains a mixture of procedures that may have different purposes: therapeutic procedures in clinical research (e.g., drugs, surgery, psychological interventions) are administered with a "therapeutic warrant" (there is reasonable evidence to support the belief that they may benefit the subject), whereas nontherapeutic procedures (e.g., added blood tests, imaging procedures, questionnaires) are administered without therapeutic warrant and solely to answer the scientific question at hand. Separate moral calculi have to be used for each of these procedures; no catchall approach to risk analysis will lead us to sensible conclusions about benefit-harm analysis.

What are these standards for therapeutic and nontherapeutic procedures? Therapeutic procedures must fulfill the standard of clinical equipoise. Fundamentally, it builds on the recognition that physician-researchers do not stop being physicians when they start being researchers. Therefore, they owe a duty of care to the patient-subject. Therapeutic procedures in the various treatment arms of a trial must be consistent with "competent" medical care, not the "best available" medical care. Freedman put this formally, as we heard earlier, saying that this tenet requires that there exist at the beginning of a trial a state of honest professional disagreement in the community of expert practitioners as to the preferred treatment.

As Professor Levine pointed out earlier, there are all kinds of equipoises lurking out there, waiting to ensnare us. If one is speaking of Freedman's clinical equipoise, the opinions of patients are quite irrelevant. Under Freedman's view, patient views are captured under something called consent. What follows from clinical equipoise is that

when there exists no effective treatment, a no-treatment or placebo control is acceptable; when there exists effective treatment, a trial ought to have an active control.

What does clinical equipoise imply for IRBs? In making this determination with regard to therapeutic procedures, it is important to understand that the IRB does not survey practitioners; they do not measure the disagreement in the community. Clinical equipoise requires that the IRB scrutinize the study justification, that it reviews the relevant literature, and, when required, that it consults with independent clinical experts. Clinical equipoise is satisfied if, and only if, the IRB concludes that the evidence supporting the various therapeutic procedures is sufficient that, were it widely known, expert clinicians would disagree as to the preferred treatment. This finding is a matter of judgment, which is why IRBs are committees with diverse expertise and experience.

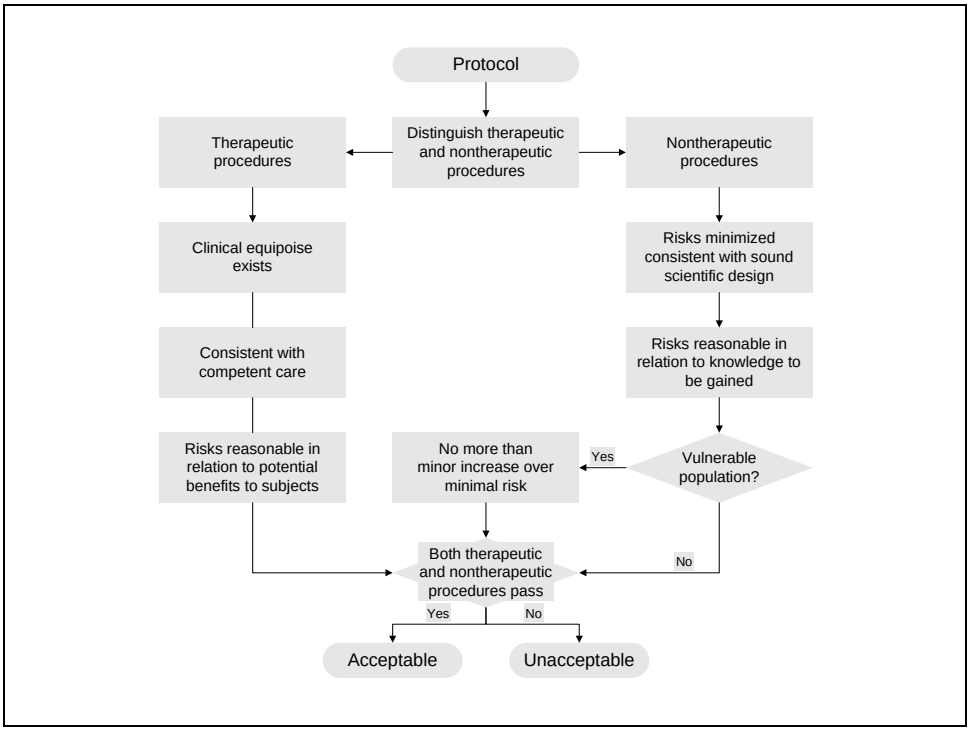
Nontherapeutic procedures are subject to different norms because they offer no benefit to participants, and hence a harm-benefit test is inappropriate. These procedures must fulfill two different moral rules: that risks associated with them must be minimized and that risks must be reasonable in relation to the knowledge to be gained. This is a harm-knowledge test. Because of the nature of these two rules logically, there is no limit to nontherapeutic risk to which a competent adult subject may be exposed so long as the question is important enough.

How does an IRB determine that these tests are satisfied? It ensures that nontherapeutic risks are minimized by, where feasible, requiring the substitution of "procedures already being performed on the subject for diagnostic and treatment purposes" (from the Common Rule). The IRB's determination that the risks of nontherapeutic procedures are reasonable in relation to the potential knowledge gain requires that it judge the study's scientific value sufficient to justify the risks to participants, which is irreducibly a judgment. Since this judgment involves, in part, an appraisal of the social value of the question, this is part of the reason why IRB membership needs to include community representatives.

Sometimes studies involve vulnerable populations and require additional protections. Vulnerable populations include pregnant women, prisoners, children, and incapable adults. There is not yet a Subpart for incapable adults, although the National Commission did recommend it in the 1970s and then the President's Commission recommended it again twice in the 1980s; we are hoping that, one day, a Subpart will afford equal protections to incapable adults. There are basically three rules with regard to vulnerable populations: they may not be included in research as a population of mere convenience, those who cannot speak for themselves are spoken for by a proxy decisionmaker, and the threshold for acceptable nontherapeutic risks are a minor increase above minimal risk. This is explicit for research involving children and it ought to be explicit for research involving incapable adults.

How do IRBs do that? The threshold for nontherapeutic risk is a minor increase over minimal risk. What does that mean? Minimal risk means the risk ordinarily encountered in daily life. In making this determination, the IRB reasons by analogy. It asks whether the risks posed by nontherapeutic procedures are in fact the same as those

encountered in daily life – such as going to the doctor and getting a physical exam or getting blood pressure checked – or whether they are sufficiently similar to these risks to be counted as being qualitatively in the same category – such as ultrasound.



For those of you more visually inclined, here is what I just said; thanks to Dr. Chad Heilig at the CDC for producing this diagram. To summarize, component analysis involves the following steps: First, therapeutic and nontherapeutic procedures in a study must be distinguished. Second, therapeutic procedures must pass the test of clinical equipoise, which means those procedures need to be consistent with competent care. Third, nontherapeutic procedures have to pass two tests and possibly three: risks must be minimized consistent with sound scientific design and risks need to be reasonable in relation to the knowledge to be gained. If there is a vulnerable population involved, then the IRB further needs to ensure that there is no more than a minor increase over minimal risk. Fourth, only if ethical requirements for both therapeutic and non-therapeutic procedures are fulfilled can the IRB determine that the benefits to participants outweigh the harms. That is a comprehensive and systematic approach to the ethical analysis of benefits and harms in research; as far as I know, it is the only such approach.

Component analysis provides clear criteria for IRBs to use in judging whether the risks of research are reasonable in terms of what might be gained by the individual or by society. Further, it allows for principled resolution of moral issues that turn on evaluation of harms and benefits in research. Finally, it provides a basis for productive future debate on these issues.

I will now turn my attention to some questions about ethics and trial design, which flow from some of these concepts, particularly the concept of clinical equipoise. The first part of the definition of equipoise is the one most commonly cited: namely, at the start of the trial, there must exist a state of honest professional disagreement in the community of expert practitioners as to the preferred treatment. In his original paper in the *New England Journal of Medicine* (1987; 317: 141-145), Freedman goes on to say something interesting that almost no one talks about: "The trial must be designed in such a way as to make it reasonable to expect that, if it is successfully completed, clinical equipoise will be disturbed." In other words, the result of a successful trial should be convincing enough to resolve the dispute among clinicians. This is interesting because it casts the ethical preconditions of clinical research as an issue in medical knowledge, which is reflected by practice based on a foundation of evidence. According to Freedman the purpose of RCTs is, ultimately at least, to provide evidence sufficient to change practice.

It follows that the ethics and epistemology or science of RCTs are not separable issues but are deeply and inextricably intertwined. That is not the way ethicists talk. Ethicists spend a great deal of time talking about informed consent. We spend some time defining impermissible avenues of inquiry (e.g., the use of placebos in RCTs in major depression) and issues like protecting vulnerable participants. All of these topics are important, but clinical equipoise suggests that aspects of trial design are matters of both scientific and ethical concern – questions such as who will be studied, what will be studied, how many will be studied, and when the study will be complete.

One of the most important papers written about RCTs was written by Schwartz and Lellouch in 1967 and called "Explanatory and Pragmatic Attitudes in Therapeutic Trials." We will talk about the ARDSNet trial later, but it strikes me that this paper captures exactly what was going on in the ARDSNet dispute – fundamentally, there was a difference of philosophies of trial design. ARDSNet investigators believed that an explanatory trial was the right thing to do and some NIH investigators believed that a pragmatic trial was the right thing to do. That disagreement led to the history of which we are all aware. It is useful, therefore, to make this difference in philosophies explicit rather than implicit, so that when these disputes occur we can recognize them as disputes in the philosophy of trial design. These two philosophies of trial design each have pervasive implications for defining the research question, study arms, eligibility criteria, and study outcomes. These philosophies are idealized as explanatory trials and pragmatic trials, but in reality they exist along a continuum.

An explanatory trial is one that attempts to discover whether a difference exists between two treatments that are specified by strict definitions. The aim of such trials is to deepen the understanding of a medical intervention under tightly controlled circumstances, akin to those found in a laboratory. The example that they give is one of a radiosensitizing drug given for 30 days prior to the first course of radiotherapy for a cancer patient. The explanatory question is whether the drug has the biological effect claimed. That would lead to an explanatory trial design in which participants would be randomized to treatment with the drug followed by radiation or to a second arm of no

treatment for 30 days followed by radiation. It is that design that gets at the explanatory question of whether the drug has the biological effect claimed.

A pragmatic trial compares two treatments under the conditions in which they would be applied in practice. It seeks to answer the question of which of the two treatments we should prefer, and therefore it is not primarily directed toward understanding but toward making a decision as to which treatment is to be preferred under clinical circumstances. Referring to our example of the radiosensitizing drug, the pragmatic question would be which treatment should be preferred in the clinic. That involves randomizing the following treatments to patients: treatment with the drug followed by radiation versus immediate radiotherapy. That is the pragmatic design.

Benjamin Freedman said the following about clinical equipoise and these two approaches to trial design:

> *"Explanatory trials purchase scientific manageability at the expense of an inability to apply the results to the messy conditions of clinical practice. Overly explanatory trials designed to resolve some theoretical question fail to satisfy the second requirement of clinical research (i.e., changing practice) since the special conditions of the trial will render it useless for influencing clinical decisions, even if it is successfully completed."*

The point here is that, in general, clinical equipoise will tend to favor more pragmatic approaches to randomized controlled trials. However, I think the point is overstated. For instance, explanatory trials do occasionally have immediate pragmatic implications, namely when they are convincingly negative, because the purported treatment has failed under optimal circumstances and therefore is exceptionally unlikely to work in the messy circumstances of clinical practice. This distinction that Freedman tries to uphold does not hold up in all cases. Nonetheless, there is a prima facie preference for pragmatic Phase III clinical trial designs. Explanatory approaches may be used, but with justification.

What about usual medical care? I would have put more distinctions in this talk if I had written it after the first couple of speakers, because those speakers made useful distinctions between protocolized care and a more as-it-is usual care arm in which practitioners do whatever they would do normally. Without making those important distinctions, I will say that, for two-arm trials, clinical equipoise would suggest a general preference for a standard medical care comparator arm in a Phase III clinical trial, where a standard, effective medical treatment exists. This design would be a pragmatic design, leaning more toward usual care circumstances; however, justifications can be invoked for using an explanatory design even in these circumstances. The first justification is that there is an explicit plan to follow an explanatory trial with a pragmatic trial, which does not happen often enough.

There might be other justifications for using this design – complexity of the treatment regimen, heterogeneity of practice, and dynamic practice standards – that will pull people in two directions. The question is how people will react to complexity.

Physicians and trial designers react in two different ways, which serves as a litmus test to divide people who are truly of an explanatory orientation from those who are truly of a pragmatic orientation. Under these circumstances of complexity, heterogeneity, and dynamics, people will polarize on the design issues. In the face of complexity, heterogeneity, and a rapidly moving standard of care, it is difficult to say what the correct question should be. An approach that says "what we need to do here is to simplify and come up with a tight explanatory design" is a reasonable approach to complexity, recognizing that it is perhaps equally reasonable for people to believe that complexity must be captured in its fullest and, therefore, a pragmatic design is the reasonable approach. In these circumstances, people will disagree about the appropriate trial design. In my view, neither approach is unethical. Although this situation calls for a scientific debate about what question is appropriate, sometimes that debate may simply be unresolvable.

Having said this, in this context I do not understand the proposal to resort to three-arm trials at all. There has been a suggestion (which came up in ARDSNet) that, in explanatory trials, those that sample from extremes may represent the extremes of practice. It is possible that each of the extremes is harmful and that usual medical care, which rests somewhere in the middle, may be beneficial; in an explanatory trial you will miss that. What is being postulated is a U-shaped dose-response curve. In the absence of convincing evidence that that is likely to be the case, we should recognize that this type of curve could be said of any clinical trial. This is what philosophers of science would call the problem of under-determination, and it applies to all science and no less to RCTs – that the results of any trial are consistent with many conclusions. The data under-determine what conclusion we should draw. Adding more arms to a trial will not solve this problem. We could add a third arm to ARDSNet, but perhaps there is a "W-shaped" dose-response curve. Applying that standard means that all clinical trials are uninterpretable, so that is not a useful approach. If there is evidence of a non-linear dose response, we should consider a dose-response design that looks at a series of levels of dose rather than using a usual care arm.

In trying to think of potential benefits of a third arm with usual care, I could not think of too many. One is that it allows us to compare treatment arms to the usual medical care arm; it certainly allows us to track changes in medical practice. But these potential benefits come with a longer list of costs. Welding a pragmatic arm to an explanatory trial is an attempt to meld potentially conflicting trial philosophies. We should step back and ask what the appropriate question is and let the design flow from that, rather than trying to impose hybrid models that aim to achieve too much. Usual care arms are generally pointless without formal hypothesis testing, and plans are needed for a formal comparison between the usual treatment arm and the other two arms in the trial. If you are going to do that, you will need a larger sample size or your power for the various comparisons will suffer. Three-arm trials are more complex and more difficult to conduct, and therefore they are more expensive.

It is not accidental that the most contentious issues in the ethics of research involve questions of acceptable benefit and harm. It is not accidental because there is not a widespread understanding of a systematic and comprehensive approach to harm-

benefit analysis. That leads to something that is all too common in the ethics of human experimentation literature – *ad hoc* ethical analysis. Component analysis provides just such an approach, and researchers, IRB members, and OHRP would find it instructive as a framework within which to systematically and consistently interpret Federal regulation. This approach and clinical equipoise (one of its key concepts) suggest a preference for pragmatic Phase III trial design, although it is fair to say that explanatory design may be used with justification.

Discussion

Q: In discussions of explanatory versus pragmatic trials, intent to treat analysis usually comes up. Even if a trial is designed as an explanatory trial, the treatments are not necessarily given as you want them to be given. Adding an intent-to-treat analysis to an explanatory trial has the advantage of being much more real-world and related to clinical practice, but also you can define what the treatments were.

Dr. Weijer: That came out really clearly this morning. Intention to treat is an important part of trial design. One thing that came out of the first talk this morning was just how nebulous the concept of usual care is. It is fine to say we will have a usual care control arm, not a protocolized control arm, but that will be driven by institutional and regional events.

Q: I would like to extend your ethical underpinning of research by asking if you would agree that there is an implicit agreement between the subject who signs informed consent and the researcher – that the researcher will deliver a reliable result in exchange for the subject's participation. I find it difficult to imagine how a thinking person would agree to engage in a clinical trial, aside from the therapeutic misconception, if that person were not convinced that the trial would produce reasonable data.

Dr. Weijer: The short answer is, yes I do agree although I would not cast it quite in those terms. I would say that validity – the reasonable expectation that the trial will produce an answer to the question asked – is both a scientific and ethical prerequisite.

Q: Agreeing at least generally that there is an ethical imperative, I am puzzled by (1) the emphasis on using a "wild-type" control group or usual care, which makes interpretation difficult; (2) the willingness to engage in pragmatic (effectiveness) trials before efficacy data has been established, which makes interpretation of the negative result extremely difficult if possible at all; and (3) the willingness of the clinical research community to employ rescue therapy – a more appropriate term would be "desperation therapy." (But if you couched it in that more appropriate term, no parent would agree to engage that therapy, which involves things like crossover or use of the active agent in the control group or other things that subvert the scientific validity of the trial.) I am pleased to hear you make a good case for an ethical imperative to do good clinical research, independent of the scientific imperative to do good clinical research.

Dr. Weijer: I will add a caveat – to point out the primacy of Helsinki V. The interests of the subjects must ALWAYS come before the interests of science. There may be various scientific designs that appear to offer greater internal validity, but those designs may be impermissible for ethical reasons. I have a particular type of study in mind, but we are not going to go there!

Dr. Levine: The use of the term "rescue therapy" earlier this morning was intended to mean something different from your interpretation, and the use of the term was different from the customary use. Rescue therapy is a term often used to talk about something that is done to salvage someone from a threatening complication. I first heard it used in terms of leucovorin rescue during treatment with a folate antagonist. The context I heard this morning was in a placebo-controlled trial in symptomatic therapy, not with a life-threatening situation. For example, if you are doing a placebo-controlled trial of a new analgesic in the treatment of headaches, you can tell the subject he/she is free to withdraw at any time. We request that such a subject tell us why they are withdrawing, because in case the reason is that the individual cannot tolerate the headache any longer, the researchers will provide a known effective analgesic. So we are not talking about something that has quite as threatening implications.

Dr. Goodman: I would be interested in your perspective on this knottiest ethical issue – relevance of background conditions. This often explains the gap between standard of care and usual care. Costs loom large. The reason many practitioners might not have used that therapy is that they did not think the cost of the therapy merited its possibly minor benefits. Cost is often thought to be an ethically unacceptable reason not to use an efficacious therapy. There is a whole spectrum of reasons that may or may not be ethically relevant. It may be that clinicians think there is a side effect that does not in fact exist or clinicians may think that a side effect is ethically or clinically irrelevant. Can we take advantage of these large regions where possibly effective therapies are not used for legitimate (or possibly illegitimate) reasons and include them as part of the usual care package? Can we take advantage of that fact that might be called competent medical care but, by the standards of evidence, may or may not be appropriate care?

Dr. Weijer: That is an exceptionally difficult question. Pragmatic trials are all about capturing background conditions. If the purpose of a trial according to clinical equipoise is to change practice, then the first thing to do in designing a trial is to figure out what is required to change practice. If it is issues of cost, dosing, or intensity of monitoring, then those issues must be addressed. It will not be useful to have evidence that a treatment works but it can be afforded by no one. Cost is a relevant ethical consideration and I have defended the use of placebo-controlled trials, much to my regret, in circumstances in which the proven-effective treatment is not affordable.

Then you ask an even more difficult question: Can we take advantage of pockets of variation in care? There is a common misconception that the ethics of research demand that we provide optimal medical care for subjects or the best available medical care anywhere. That is false; that cannot be true because competent practitioners do disagree and vary in their practice. That is why I quite purposefully went to a lower

standard, which is “competent medical care.” If we use the standard of competent medical care, then we still achieve the ethical end of ensuring that subjects are not medically disadvantaged in a meaningful sense by virtue of enrollment in a clinical trial. What I would not want to allow is the use of pockets of poverty that exist in this country as an excuse to do placebo-controlled trials because no one can afford treatments that we know in a local sense to be part of competent medical care. I would want to draw a line somewhere, but it will be tricky to recognize where that is.

Dr. Levine: As I recall, your willingness to consider use of placebo controls in circumstances in which there was a known effective therapy was sharply circumscribed, where the purpose of the research was to develop a therapy that was affordable in the population in which you did the study. There were plans in place to make the therapy, if it turned out to be effective, available to that population.

Dr. Fleischman: You did not use clinical equipoise to justify that placebo-controlled international trial. The clinicians in those countries, who were experts, knew full well what the appropriate therapies were to help their patients, they just were unable to obtain them. So there was no disagreement as to what the appropriate expert opinion would be about the treatment. How do we use clinical equipoise in this way to justify that action?

Dr. Weijer: I am trying so hard not to go down a number of roads! If the treatment is known to be effective but is not available, then it is not part of reasonably competent medical care in that country. You can know in theory that an expensive medicine is effective, but if no one can afford it and the health care system does not provide it, then it is not part of competent medical care. If it were part of competent care, then every physician in that country would be guilty of negligent medical practice. We do not want to say that. The purpose of a trial is to change practice, so we should ask a clinically relevant question in the setting of a developing country. The clinically relevant question is not how does this affordable regimen compare to a completely unaffordable regimen; it is should we adopt, as a matter of health policy, an affordable regimen in light of the fact that we currently have nothing available.

Challenges of Practicing Evidence-Based Medicine: Integrating Science and Clinician Experience in Patient Care

R. Brian Haynes, M.D., Ph.D., FACP
McMaster University

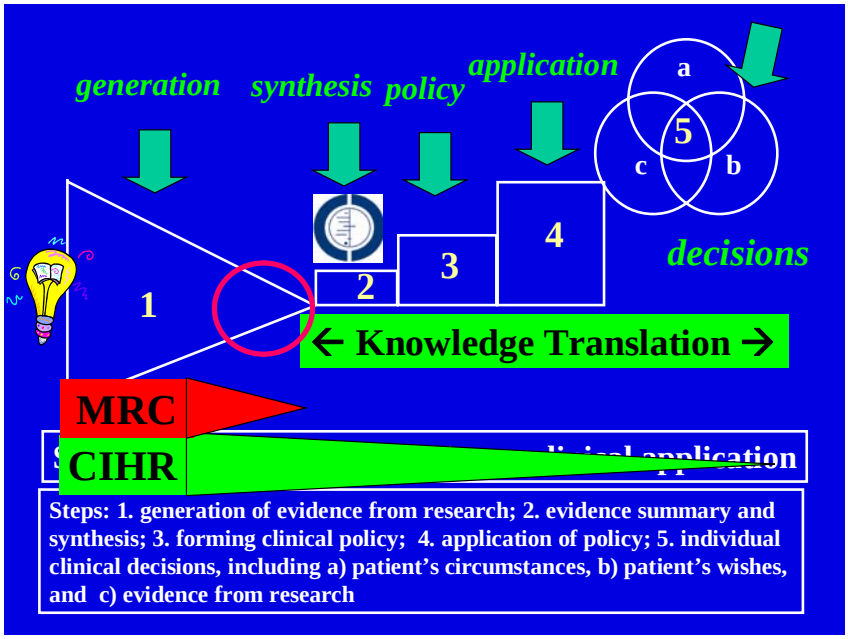
Challenges of Practicing Evidence-Based Medicine: Integrating Science and Clinician Experience in Patient Care

R. Brian Haynes, M.D., Ph.D., FACP, McMaster University

Dr. Haynes is professor of clinical epidemiology in medicine and chair of the department of clinical epidemiology and biostatistics at McMaster University in Hamilton, Ontario. He received his M.D. degree in 1971 from the University of Alberta and his Ph.D. degree in clinical epidemiology from McMaster. He joined the faculty of Health Sciences at McMaster in 1977. His main research interests are methodologies for evaluating healthcare interventions and improving healthcare through enhancing the validation, distillation, dissemination, and application of healthcare knowledge.

Clinicians are only one of the decisionmaking groups that could benefit from the knowledge that we are generating through trials; the others include policymakers, managers of healthcare payers, clinicians, patients, and the public. Many decisionmakers are on the receiving end of the knowledge we are generating; hopefully, we will be able to provide information that will be useful to all of them. The key question to which decisionmakers want an answer is: how does this new or alternative treatment compare with what is usually done? That is pushing toward the "usual care" end of the spectrum of the possibility for comparators. If we really wanted the answer to this question, we would be pushing for "N of 1" trials, in which individuals take one treatment and then the alternative treatment in a random sequence and decide which treatment is best. Such a design is impractical or unfeasible for most treatments.

My objectives are to review the steps from evidence generation to application, to provide a model for evidence-based decisionmaking, to comment on the methods of usual care trials, and to underscore the importance of usual care comparators in reaching satisfactory patient-centered, evidence-based clinical decisions.



This slide traces the steps from evidence generation through application. On the generation end, someone gets a bright idea and throws it into the testing system. We have a robust biomedical research enterprise that tests these ideas. Most of these ideas do not hold up and are thrown out; that is a good thing because we do not have to live with the ideas just because they are interesting.

Only a small trickle of these ideas makes it through all stages of testing to come out to the next step, which is synthesis – putting the evidence together. At this point we put together all the trials on a given topic and try to get a signal out of them. Very few individual investigations give a compelling message, and even the compelling trials do not cover all the range of possibilities for intervention. Synthesis is an essential step if we are going to move the information along to application.

The next step is policy. Once we have the clear signal from research, it is not a no-brainer decision what to do with it. In fact, we need to take that evidence and be respectful of its strengths and limitations, on the one hand, and take into account the circumstances and the resources we have for applying that evidence, on the other hand. One of the explanations for small-area variation, for example, is that the resources are different in different locations. Even with the same evidence, the decision about what to do may be substantially different from location to location and this will be for circumstances that are far from sinister but are just practical reality.

Once we have decided what we want to do, we have to figure out how to do it. That is the application step, which is a major step for many types of innovations. Some of the innovations are simple to apply – they exchange one product for another of the same sort – but most require reallocation of resources from one form of management to another, including training for the staff as they will have to gain experience with the new product even if it is similar to previous treatments.

After all those steps, we get to the decisionmaking step. I will not go into this in detail, but it is a complex process of taking into account the evidence, patient circumstances, severity, and patient wishes to arrive at a negotiated plan of action.

There is a long path here. The path from original discovery through application is now under the rubric of knowledge translation, a relatively new term that describes not just a pathway from evidence generation to application but also a research agenda. Many of these steps are not well understood; it will require decades of new research to understand them and to make them work better than they do at present.

The Medical Research Council in Canada did not fund any knowledge translation research. Its successor, the Canadian Institutes for Health Research, has expanded the scope of what it will fund but its first competition in Canada for knowledge translation was just in 2002 and that competition was for a limited budget. CIHR has now opened up the gates for knowledge translation research because it is under pressure from politicians to show that the money spent on health research really does do the public some good. To me, that is a very important realization about which the Canadian funding agencies are finally getting serious.

We will now look at this from the perspective of explanatory/efficacy trials on the early end of clinical testing, and management/effectiveness trials or pragmatic trials on the other, advanced end. Most of the trials are at the explanatory/efficacy end; at the management end, we get many fewer trials. We get so few management trials that a big gap is left between the new knowledge generation and the application steps. For drugs, for example, most of the trials are at the explanatory/efficacy end of the spectrum and we never get to the management trials, so if people are trying to decide what to do with the information, they have to fill in the gap between what the trials show and the next steps, which could lead to application. They are going to have to make it up, because the information is simply not there from the trials and product testing legislation does not require it before licensing for general use.

That is true for drugs, and the situation is much worse for devices and services. Here we do not require randomized trials, at all, let alone trials that are at the management end of the spectrum. For a lot of devices, we have to leap a huge information gap before application. As a result of that, we indulge in speculation and we adopt a lot of things that do not work. That is a pity and a shame. We should think about an FDA for devices, not just for drugs.

Models are simplistic, of necessity. They cannot possibly take into account all of the things that would take place in a complex process. For evidence-based clinical decisions, the slide depicts the main elements that we think are important. They begin with the clinical circumstances for patients. Obviously, we need to define the problem, make a diagnosis, decide how severe it is, and determine whether the patient has competing comorbidities that may complicate their treatment, whether they have allergies, or whether they cannot afford the treatments available. Only after that intense clinical process can we consider what research evidence might be brought to bear on which treatments should be offered – the alternatives and how they fit with this person's clinical circumstances. Then we offer the alternatives to the patient and ask the patient what their preferences are and to consider what their actions might be. I work in a diabetes clinic. Patients there would like not to take medications, they would like to be able to treat their Type II diabetes with weight reduction and exercise, but their track records show that they are not getting to the goals of treatment on this basis. That must be taken into account as to how strongly one recommends alternatives. That is where clinical expertise comes in, and expertise is in fact required at each of these steps if we are to come to an appropriate, balanced decision about what alternatives to offer and how to help patients make a decision about them and implement them.

That is the basic model for clinical decisions. Research evidence is clearly only one part of the model. In fact, in many situations, it is not even close to becoming the determinant part of it. Fortunately, we have increasing amounts of research evidence that are making it possible to make sounder decisions for patients because of the available evidence. Unfortunately, at present, sound evidence from research does not cover even a large fraction of much of medical practice.

The basic questions answered by trials are: can treatment X work under ideal circumstances – efficacy or explanatory trials – and does treatment X work under usual

circumstances – effectiveness or management trials. The first of those questions is the one that most trials try to address. The purpose of this conference is to discuss "compared to what" – what is the comparator for this? This is a spectrum; trials are not at one extreme or the other. The possible comparators include placebo, best care, or usual care. In fact, there are no trials that are perfectly at the explanatory end; we cannot create perfectly ideal circumstances and there is no such thing as a perfect management trial because we have to invite people to join the investigation and we have to get their consent to do so. As a result, we cannot include in a trial everybody who might be in that trial.

The possible valid comparators include index treatment (the new treatment being studied) plus best existing treatment, versus placebo (or nothing) plus best treatment. I want to show you the limitations that can be imposed on the decisionmakers if we have this type of comparison. As an example, I will use the North American Symptomatic Carotid Endarterectomy Trial because I was intimately involved in it through the methodologic coordinating center at McMaster University. (The University of Western Ontario was the lead center for this trial.) This trial compared carotid endarterectomy with no placebo, because there was no appropriate placebo, and both groups got best medical care. The question that was answered by this trial was:

> *"For willing patients with recent hemispheric transient ischemic attack (TIA) and partial stroke, who had ipsilateral internal carotid artery stenosis of 70 percent to 99 percent (so they had to have angiography to be included in the trial) and no greater distal stenosis and low operative risk, carotid endarterectomy plus best medical treatment by highly competent surgeons is beneficial compared with best treatment alone."*

It was appropriate to answer this question with a trial because, before the trial, we had not had a randomized trial of carotid endarterectomy, and that was all the surgeons would agree to in terms of being able to mount a trial. So we would not have had an answer to any question if we had not picked this particular question. In practice, what happens with the type of answer we generated is that we can hardly expect practitioners to know what to do with the answer. The patients were highly selected, as were the surgeons; all patients had angiography, an invasive procedure; and research funds were provided to ensure that all patients were closely followed and received best medical care. The circumstances of usual clinical practice are not so pristine. My estimate according to current evidence is that, in usual practice settings, only about 1 in 10 people who could benefit from this procedure actually receive it. Further, among those people who are provided with the procedure, only about 1 in 4 meet the stringent criteria for this trial. There is a huge mismatch between what this study showed and what patients are offered in practice.

From a knowledge translation perspective it is also possible, and better, in trials to compare index treatment plus usual treatment versus usual treatment, and we have some cardiac defibrillator trials that are on that model and that provide useful information. We can also compare index treatment alone with usual treatment. Most health services trials are of this nature – a randomized trial with a complex intervention

and a usual-care comparator. If these comparators reflect what can only be done in trials, they are found in the left hand side of the testing spectrum, providing not so much practical/useful information. This is the problem I hope this conference will address: how to push trials toward providing more practical and useful information. The trials become more useful as they push toward the right end of the testing spectrum.

Regulatory requirements for drugs – not true for devices and services – are mainly at the placebo-controlled comparator end. It costs more to test a new intervention if we go to a "current best-care" comparator model, but it costs not as much to use a usual care comparator as it does to use a best care comparator. The problem is not just money; the resources required to do a best care comparator are substantial – you have to select the patients, you have to select the treatments, and you have to provide extra support services to get them implemented.

One solution is large simple trials in usual-care settings. These have been pioneered through the United Kingdom and the Oxford Clinical Trials Group in particular. They involve thousands of patients in many locations in usual care settings, with a minimalist data collection set. The ideal for this is the amount of data can be put on the back of a postcard. Huge infrastructure is not required for these types of studies and they allow factorial designs, with the testing of several interventions and their logical combinations. The HOPE trial is an example here. Published in 2000, it is a 2x2 factorial trial of ramipril versus placebo and vitamin E versus placebo for lowering cardiovascular risk. There were 10,000 patients, 267 centers, 4.5 years of followup, and a definitive answer: ramipril works and vitamin E does not work.

Unfortunately, the comparator is only one of the difficulties inhibiting practical application. We have problems with patient selection, practitioner selection, supportive add-on treatments, and patient monitoring. For example, in an efficacy or explanatory trial, we typically will select patients who are high risk and highly likely to respond to the treatment. We get providers who will be sure to follow the treatment protocol, we optimize the index treatment (treatment being studied), and we often use a placebo comparator or a not-so-strong comparator to maximize the difference in effects between the intervention and control groups. We optimize the support for patients and we have obtrusive monitoring, in which the patients are seen much more frequently than they would be in usual practice. They are scrutinized so that they know that their contribution is important and so that they continue with followup.

Decisionmakers want something that resembles usual practice circumstances, so they would like to see more representative patients included in these trials, and they would like to see how treatments actually work under usual circumstances. They do not want a situation in which there is a lot of care that they cannot afford to provide for patients in the followup period. Decisionmakers, both practitioners and patients, also want more information than is typically provided about adverse effects. They want information about quality of life, cost, cost effectiveness, predictors of who will benefit, and predictors of who will be harmed. These are all potential benefits of large simple trials, because we can define, in advance, the subgroups in which we are most interested and

power the studies adequately for subgroup analyses that will provide us with information about, for example, which patients will benefit and which patients will not.

Here is an example of the type of information that is useful in practice that typically is not produced by clinical trials. If you have nonvalvular atrial fibrillation, the chance is about 5 percent that you will have a stroke in the following year, but that chance is not the same for every patient. The age of the patient makes a difference and the history of hypertension, diabetes, and prior stroke all increase the risk that you will get a stroke from your atrial fibrillation. We can tease these out quite well, so that a person under age 65 with none of these risk factors has only a 1 percent chance, and the treatment cannot realistically be expected to benefit a person with a risk that low, especially when considering the adverse effects of the treatment. On the other hand, other patients have a more substantial risk and they are much more likely to benefit from this treatment.

We can also take a look at adverse effects from the intervention, Warfarin, which is the recommended best treatment. Age, history of stroke, gastrointestinal bleeding, and comorbid conditions all affect the likelihood of harm from Warfarin. This risk rises substantially if you have a number of these factors. It is possible to define a patient population that is at high risk for having a stroke and is at relatively low risk of suffering the complications, thus making a much better match between knowledge of the benefit of this treatment and who should receive it and under what circumstances. That is the type of information we need, and we could get it from trials if they were properly designed to ask those questions.

Do we need the trials, or can we rely on physician judgment? In this case and in many situations, we cannot rely on physician judgment. Their experience is not adequate to define the risks, in this instance, of how likely a person is to bleed if they take Warfarin. I could show you more interesting results in which physicians are all over the map in terms of their estimates of risk of harm from Warfarin; there is no standardization in medical practice about how these patients are treated, which is a major cause of practice variation.

Decisionmakers want answers to these questions and I hope that this meeting can push us all toward recommendations for more usual care trials, so that practitioners, patients, policymakers, and managers can make better decisions about what to do with the information coming from research.

Discussion

Q: Should effectiveness trials follow efficacy trials?

Dr. Haynes: There is every justification for doing efficacy trials first to determine whether something can work. If it cannot work, there is no point in doing the expensive effectiveness trials, and it could be argued that it is unethical to do those trials. I do think there is a sequence. Only in something that has extremely low risk and high potential for benefit would you want to do the management trial first.

Q: [off camera]

Dr. Haynes: From a practical perspective, there is information about which patients are likely to come to harm and which would benefit from Warfarin, but those predictors are not as good as they could be. If we have genetic information concerning responsiveness to treatments and likelihood of adverse effects, it would strengthen greatly what we get out of application in the community. This can be part of each clinical trial. If we collect information or just store blood samples until we understand the genes better, we can then look back at the information and find out how well we can define who is likely to benefit and who is likely to come to harm. I agree that that information could be valuable. It does not have to come from randomized trials, but the best time to collect it is in randomized trials when we have excellent followup data to be able to use it.

Q: Please comment on the paradox that the explanatory trials would have narrow confidence intervals (but would not be widely applicable) whereas the management trials, because of the heterogeneity of people involved in the investigation, would be expected to have wide confidence intervals (then you would not know how to apply the results because of that situation).

Dr. Haynes: No one trial will answer every question. However, with the large simple trials you can get narrow confidence intervals on how things work within disease strata if you include enough patients. But who could afford to include lots of patients? However, many of these large, simple trials have been run on substantially smaller budgets than some of the efficacy trials. They have a much smaller infrastructure. When we did the NASA trial, we had 30- to 40-page case report forms that required that we hire professional staff to collect the information; there is no way we could make the practitioners do it. That kind of data collection is very expensive. If we off-load the data collection part of it, we can actually increase the number of patients and the number of subgroups of patients for which we can define relatively precise estimates. I do not think it is a cost issue so much as a will issue, in terms of managing these trials.

Q: Please comment on the quality of data from a large simple trial compared with a situation in which you have much better control over the collection of data.

Dr. Haynes: If a large simple trial has simple outcome measures that are clinically important, for example life or death, then the complexity of data collection is not huge. If you get into more complicated data collection, which may be warranted at the efficacy stage in which you are trying to look at mechanisms at the same time as you look at treatment outcomes, you will have to spend more on data collection. The large simple trials do not necessarily require complicated-enough data collection to make it unfeasible to obtain good quality data. Followup is the key aspect of these trials, to make sure the patients are followed over time to understand what happened to them.

Q: [off camera]

Dr. Haynes: I do not know that there is a perfect example. As I recall the HOPE trial, the effect of Ramipril was independent of blood pressure lowering; it was not a result of blood pressure lowering. I have not looked at the trial recently to look at the nuances of how the results can be interpreted. I will not defend the HOPE trial any more than that. I just used an exemplar of a different approach to data collection and management, involving lots of people in the investigation. ISIS trials, for example, had multiple treatment groups with multiple factorial design and came to conclusions that were robust because they were looking at life and death as outcomes, not because there were data collection procedures that teased out nuances of the data.

Q: It is normal to characterize the robustness of the scientific research in terms of its effectiveness in the field as part of the science itself. The idea that these are fundamentally two kinds of trials is not universally applicable across science and the application of scientific knowledge. To analyze different medications and different problems to determine the robustness of the data and the likelihood that the scientific results can be translated in the field is a core part of the scientific enterprise, particularly when dealing with the usual care controls. Robustness is a critical issue in the fundamental application of the results.

Dr. Haynes: Agreed.

The Acute Respiratory Distress Syndrome Network (ARDSnet): Lessons Learned for the Design of Critical Care Research

B. Taylor Thompson, M.D.
Massachusetts General Hospital

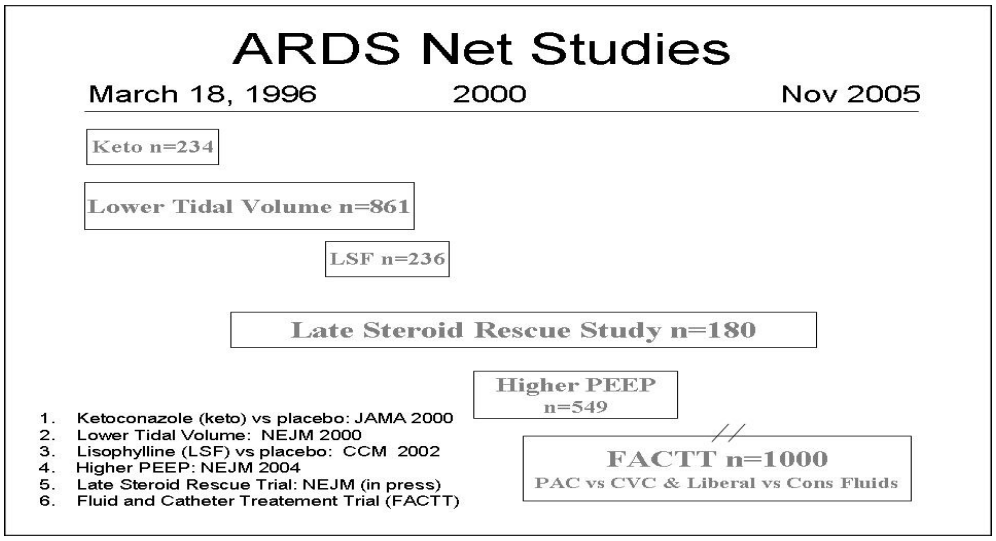
The Acute Respiratory Distress Syndrome Network (ARDSnet): Lessons Learned for the Design of Critical Care Research

B. Taylor Thompson, M.D., Massachusetts General Hospital

Dr. Thompson is director of the Medical Intensive Care Unit, medical director of the ARDS Network Clinical Coordinating Center at the Massachusetts General Hospital, and an associate professor of medicine at Harvard Medical School.

We will review the ARDS clinical trial study that was instrumental in continuing this discussion about the role of usual care in control arms. We will also discuss lessons learned from the conduct of critical care research in the last decade. The design challenges faced by the ARDSnet investigators are common to clinical inquiry in the intensive care unit generally.

The ARDS network is an NIH/NHLBI-funded group of established investigators that currently has 18 clinical sites. The network has completed 6 primary studies involving more than 2,600 patients over nearly a decade. An independent protocol review committee must approve all trials and approved trials are monitored by an independent DSMB.



Here is a schematic of the trials completed during the past 9 years. The Ketoconazole (Keto) and Lisophylline (LSF) studies were randomized placebo-controlled trials (which we are not talking about today other than to say that they were factorialized on a lower tidal volume trial) and a followup trial that examined lower tidal volume combined with higher positive end-expiratory pressure, or PEEP, which is a putative lung protective approach involving residual pressure in the thorax after exhaling. The fluid and catheter treatment trial (FACTT), another 2x2 factorial trial, examined pulmonary artery and central venous catheter effectiveness as well as the efficacy of liberal versus conservative fluid management. In the lower tidal volume trial, critics charged that the control group did not represent best current practices and extended this charge to

include the control arms of the higher PEEP and FACT trials. We have already heard that this term is problematic but that was the initial charge.

Critics also pointed out that randomization to one of these two strategies (higher or lower tidal volume and higher or lower PEEP; or fluid liberal or fluid conservative) deprives patients of individual titration of therapy based on patient care preferences and unique patient needs. If one arm is superior, it remains to be proven that this arm is superior to usual care practices and, in the parlance developed this morning, answers the efficacy or explanatory question but does not answer the effectiveness or pragmatic question.

The OHRP investigated these and other issues, ultimately determining that risks to participants were minimized and reasonable in relation to anticipated benefits and the importance of the knowledge that was expected to result. However, the OHRP went on say that they believed that the interest of future human subjects would be best served by further discussion by the scientific and bioethics communities of the issues regarding appropriate research design in the absence of a definable standard of care. From this morning's discussion so far, it is clear why the OHRP would be asking for help about this – it is not an easy question.

We would like to review two case studies in detail to highlight the challenges of clinical trial design in the absence of a well-defined standard of care. Please suspend your knowledge of the treatment results to limit the pernicious hindsight bias. We will then talk about three-arm trials with usual care as a control and the role for usual care in effectiveness trials and will review an example of best practice developed from early efficacy to subsequent effectiveness trials. We will finish with summary and recommendations.

Randomized Trials in the ICU without Usual Care Controls

	Randomized Groups
Brochard (AJRCCM 1994)	Three weaning strategies
Esteban (NEJM 1995)	Four Weaning Strategies
Hebert (NEJM 1999)	Two transfusion thresholds
ARDSnet (NEJM 2000)	Two ventilator strategies
Sandham (NEJM 2003)	PAC+ protocol Rx vs CVC
Chastre (JAMA 2003)	8 vs 15 days of Abx for ICU Pneumonia
SAFE (NEJM 2004)	Albumin vs saline resuscitation

Let us begin with an overview of a convenience sample of major ICU clinical trials (listed above). Brochard and Esteban randomized patients to various weaning strategies in the early to mid 1990s. Hebert and colleagues randomized to two transfusion thresholds. Chastra and colleagues compared two different durations of antibiotics for ICU-acquired pneumonia, and recently the SAFE investigators in Australia and New Zealand randomized patients to albumin versus saline for early fluid resuscitation. None of these trials used usual care as the control group. They tested competing

treatment strategies available in usual care practices when it was uncertain as to which was approach was superior. None of these designs answered the pragmatic question of whether "customization" to meet individual patient needs was the best approach. For example, the possibility that one weaning method may be best for a given patient, or that a trial-and-error method of various weaning strategies might be superior to random assignment to a single therapy was not tested. Similarly, the Chastra trial did not test a customized antibiotic duration, in large measure because clinicians lacked the information from randomized trials to make these treatment decisions. Such pragmatic trials, based on the knowledge gained in these efficacy trials would come in time, as I will discuss later. That these trials did not test customization does not necessarily mean they are unethical or problematic. In my view, equipoise was present as to which treatment was best at the time. Furthermore, the willingness of attending physicians to allow their patients to participate suggests to me that they were probably uncertain as to the best treatment option (trial-directed treatment or their own customized care) for the *individual* enrolled patient.

We will now focus on two trials: the Canadian Transfusion Study of two transfusion practices conducted by Hebert and colleagues and the ARDSnet trial of two ventilator strategies.

The risk of transfusion versus anemia is a clinical problem with which clinicians in the ICU struggle daily. We sample substantial amounts of blood for diagnosis and monitoring and the bone marrow does not produce new red blood cells adequately during critical illness. Thus, patients become anemic quickly. ICU clinicians must balance risks of transfusion – including immunosuppression and possibly immunostimulation in systemic inflammatory conditions – versus the risk of anemia, including inadequate delivery of oxygen to tissues. A survey of Canadian ICU clinicians revealed variability in practice and an RCT showed improved hospital mortality with a restrictive transfusion strategy, although the primary endpoint – 28-day mortality – was not different.

This trial turns out to be quite controversial, though not in my view. Pre-study activities included a survey of Canadian clinicians. They were asked to describe the level of hemoglobin at which they would transfuse a unit of red blood cells in four clinical scenarios. The survey revealed that some clinicians in one scenario would wait until the hemoglobin fell to 6.5 mg/dl to transfuse while others would transfuse at the recommended threshold of 10 mg/dl. The distribution of transfusion thresholds was normally distributed for this scenario around a mean of ~8 mg/dl. The province in which the clinician practiced and an academic versus community setting was associated with variation in practice. That clinicians caring for the same patient would make very different treatment decisions is an example of *unexplained practice variation*, and this example gives you a sense of the magnitude this variation.

Not all variation is unexplained. When clinicians were presented with another scenario – a critically ill patient who was actively bleeding – clinicians said they would transfuse at a higher hemoglobin threshold (anticipatory transfusion). This shift to a different practice for a different clinical problem is evidence of customization, though there was

still wide variation in practice for the bleeding scenario. Thus, the distribution of decisions for transfusion in the ICU reflects both unexplained variation (or practice style) and customization.

What did Hebert and colleagues do with this pre-trial survey information? They designed an RCT to test the risk of transfusion versus anemia and chose what I will refer to as an "A versus B" clinical trial design. "A" was a restrictive transfusion practice (transfuse for a hemoglobin <7/dl), which is not well represented in usual care practices. Accordingly, this strategy was tested in a Phase II trial that suggested safety (a new treatment that is not well represented in usual care practices should undergo Phase II investigation). Investigators picked the traditional transfusion threshold of 10 g/dl as their comparator group, and it is quite clear that, in many of these scenarios, usual care had moved off that old recommendation toward lower thresholds. In designing this trial, one has to ask the question, "Is there equipoise to randomize a *bleeding patient* to a 7 g/dl transfusion threshold?" While I cannot speak for the Canadian investigators, apparently the answer was "no." Bleeding patients were then excluded from the trial. This process ultimately identified a range of usual care practices for a number of conditions in which there was equipoise to randomize treatments and attempt to answer an important question.

Now we will talk about "A versus B." The advantage of "A versus B" has already been discussed nicely today. It will give us a crisp answer to the question; in the example we are discussing, an answer about the risks and benefits of transfusion. But the problem is that testing "A" and "B" will not help us evaluate customization. So what about an "A versus *de facto* usual care preferences"? Usual care is the full spectrum of practice. So one trial might be "A versus the full range of transfusion practices." Some have argued for this approach, which would be a pure pragmatic trial. The problem is that *de facto* usual care is variable and difficult to describe, making it difficult to draw inferences as to the reason that usual care may be better or worse than "A". Furthermore, usual care may have extreme practices in it, which might create ethical concerns for clinical trialists who would not want to randomize to usual care under such circumstances. Clinicians may refute a trial where "usual care" is found to be superior to "A" by saying, "They compared restrictive transfusion to what clinicians are doing in Hamilton or Montreal; that's not my practice."

An alternative would be to randomize to a restricted domain of usual care called "competent care." This may solve the problem of the outlier effect and results in a comparison of "A" versus "competent care," a reasonable effectiveness trial design. The problem is, if competent care is better than "A" and if competent care encompasses a range of practice styles (likely), it would be unclear as to which aspect of competent care should now be adopted more uniformly in current care. As Claude Bernard noted: "When you don't know what you're looking for, you won't know what you've found." That is one of the major problems with usual care as a control group – the inability to draw generalizable conclusions and mechanistic inferences.

Hebert and colleagues' "A versus B" trial provided valuable information to improve patient care and stimulated research into the possible causes of morbidity associated

with red cell transfusion. It did not answer the customization (effectiveness) question. I give you this example so all of you can consider which study design would have been most helpful to you in changing your practice. There is no correct answer.

The second example is the ARDS network trial, which tested a traditional approach to mechanical ventilation compared to a lower tidal volume approach. One of the challenges with this clinical trial is that usual care practice was changing during the design of this trial.

Traditionally recommended tidal volumes were between 10 and 15 ml/kg measured weight. Some experts were recommending tidal volumes of more than 20 mg/kg. However, an exciting body of animal research has long suggested that lungs stretched from a ventilator-delivered tidal volume – the size of the delivered breath – might lead to lung injury. This so-called ventilator-induced lung injury could worsen ARDS or delay recovery. In 1993, consensus statements based largely on animal studies suggested that clinicians limit tidal volume to avoid plateau airway pressures of greater than 30 to 35 cm of water. The consensus called for clinical trials to determine if this approach worked.

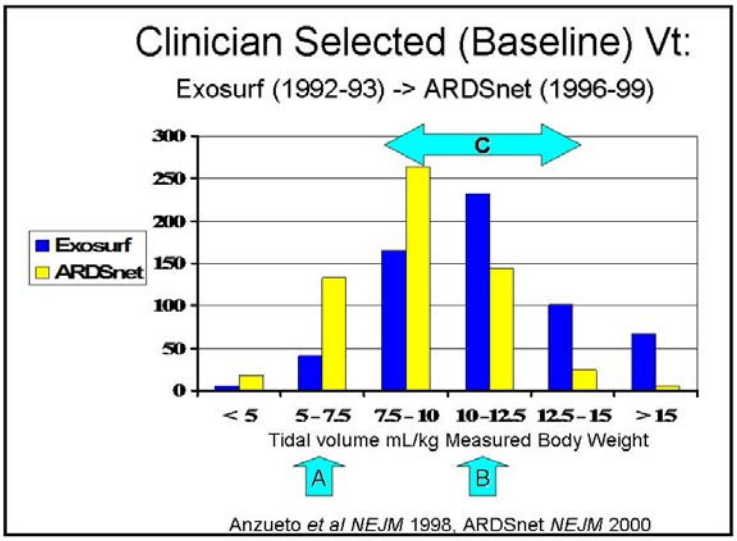
Clinicians were left in a quandary: should they continue with the traditional tidal volume approach, which puts a higher priority on gas exchange, acid/base balance, and comfort, or use the lower tidal volume approach, which reverses these priorities? Safety of the lower tidal volume approach was supported by some encouraging Phase II trials. The ARDSnet Phase III trial of a traditional versus lower tidal volume approach subsequently showed a nearly 10 percent reduction in mortality with lower tidal volumes. This landmark trial indicated that ventilator-induced lung injury (or “doctor-induced lung injury” as the ventilator simply does what is asked of it) is a major clinical problem. That a well-intended, rational, traditional approach to supportive care in the ICU was costing lives was a wake-up call for clinicians. However, critics charged that the traditional group tidal volume was too high and was not customized, and thus did not reflect “best current care.” Now we return to the crux of this issue.

How did ARDSnet investigators choose the traditional tidal volume approach? A survey of clinician preferences in 1994 (by Carmichael and colleagues) showed some clinicians were still using large tidal volumes, with the majority using tidal volumes in the 10 ml/kg to 13 ml/kg range. Fewer clinicians were using tidal volumes of 9 ml/kg or less. ARDSnet investigators obtained data from a 750-patient Exosurf trial and a sepsis trial testing ibuprofen to examine actual usual care tidal volume choices in the early 1990s. Clinicians were using a wide range of tidal volumes that averaged about 10-12 ml/kg measured weight, thus the ARDS Networks choice of a target tidal volume of ~10 ml/kg measured (12 ml/kg predicted) body weight for the traditional approach that represented the central tendency of usual care practices in the early 1990s (see Figure below). After the study had started, an international survey of clinician practices conducted by Esteban and colleagues found that clinicians continued to use a wide range of tidal volume but the mean was now around 8 ml/kg measured weight. This probably represents a secular trend toward lower tidal volumes based on consensus recommendations to reduce tidal volume. In retrospect, it appears that the ARDSnet

investigators had picked a traditional tidal volume that was consistent with practice in the early 1990s but somewhat higher than usual practice in the late 1990s when the study was conducted.

The ARDSnet investigators used a tidal volume of approximately 5 ml/kg measured (6 ml/kg predicted) for the intervention arm. Because this is outside usual care practices, a Phase II study was performed at Johns Hopkins that suggested safety and demonstrated feasibility. In addition, a trial from Brazil by Amato and colleagues of similarly low tidal volumes was supportive of this tidal volume choice.

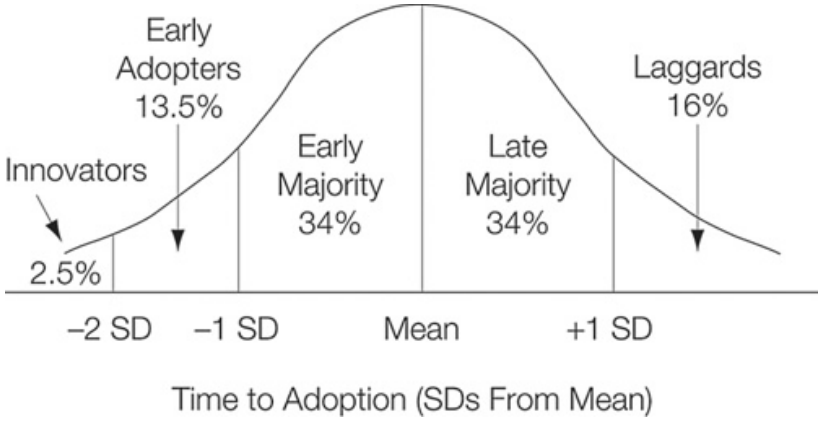
So, what might have been better control for contemporary usual care practice? The choices include randomizing to a full spectrum of usual care practices; we outlined previously some of the limitations of this approach. Perhaps restricted usual care that brackets what was considered competent care could have been used. The ARDSnet investigators chose a control group that reflected a traditional approach and also represented the central tendency of practices prior to the trial.



So why the controversy? Shown in blue is the distribution of tidal volumes in the aforementioned Exosurf trial used to design the trial. The gold bars show the usual care tidal volumes prior to study entry in ARDSnet participants from 1996 and 1999. The ARDSnet traditional group, indicated by the letter “B,” no longer represents the central tendency of usual care. Had we chosen a range of tidal volumes for the control group “C,” we would still bracket the central tendency but we certainly would not have represented as well the new range of usual care practices. Selection of B as the traditional group resulted in a controversial clinical trial because the “control” group appeared to be out of step with contemporary usual care practices, variable as they were.

What are the limitations of controlling usual care practices by either setting the tidal volume at the central tendency or protocolizing a range around the mean? First, usual

care that is not tethered to a strong evidence base is susceptible to secular change and the influence of the research environment (the Hawthorne effect). This is a real challenge for unblinded ICU trials. In our Phase II pilot study at Hopkins conducted by Roy Brower, clinicians were asked to choose a tidal volume for the control group within a range and usually chose a tidal volume that was lower than the value they had chosen prior to being approached by study personnel. They were clearly influenced by the research environment. Thus, if the control is protocolized (stabilized around a target value), it no longer "controls" for secular trends in usual care practices.



Using the above diagram from Rogers' classic text, we appear to have studied what we thought was a practice representative of the early and late majority but ended up with a control group that was "sitting" on the first standard deviation between the late majority and the laggards. Had the control group moved much further, we would have been accused even more of a "straw man trial." However, there were a number of vocal laggards and late-majority types arguing for the traditional approach. In a paper in the *New England Journal of Medicine* months before the ARDSnet trial was completed, John Weg wrote that his analysis of the Exosurf trial

> "cast substantial doubt on the view that high ventilatory pressures and volumes are harmful. This belief has led to a plethora of new ventilatory strategies, iatrogenic acid/base derangements, widespread use of prolonged muscle paralysis, and attendant morbidity. Such strategies should now be reassessed."

In my view, a compelling trial result was needed to change practice during this period and the traditional approach tested by the ARDSnet was reasonable.

A word of caution: well-reasoned physiological constructs strongly supported by preclinical data may lead us to the wrong practice changes. Many clinicians used a class of medications to suppress arrhythmias after myocardial infarction based on such evidence only to learn from a compelling randomized clinical trial (CAST) that this approach resulted in higher mortality.

What about customization of tidal volume? Approximately 96 percent of responders to the aforementioned survey by Carmichael et al. said they adjusted tidal volume based

on airway pressures; unfortunately, the literature did not describe how they did so and what thresholds were used. ARDSnet investigators examined plateau airway pressures used by other investigators and by consensus developed rules for customization of tidal volume for airway pressures; these rules were subsequently approved by the Protocol Review Committee (PRC). We learned that surveys are not terribly reliable – what doctors say they do and what they actually do may be quite different. The reality was that in only 45 percent of patients in the aforementioned international survey was a plateau pressure even measured. In Gordon Rubenfeld's recently published trial of clinician practices in the King County (Washington) lung injury project, an amalgam of private practice and academic medical services, plateau pressure was measured in only 32 percent of patients. This is akin to a doctor claiming to control blood pressure to a specified level and then neglecting to measure the pressure.

There was little or no evidence of customization of tidal volume to airway pressures in usual care for subjects prior to enrollment in the ARDSnet trial. Two groups of investigators have looked at the data and come to different conclusions about evidence for customization, a good illustration of how hard it is to determine physician intentions in variable usual care.

Other ARDSnet controversies:

- Control groups did not reflect "best current care." The ARDSnet response was that there was no best practice at the time and the traditional approach was still in use and was being defended.

- Comparisons of trials and subgroup analyses suggested that intermediate tidal volume or "customized" tidal volume was superior to either high or low tidal volume. Investigators countered that meta-analyses and subgroup analyses do not support this conclusion. A recent analysis from Hager et al. suggested that even lower tidal volumes might be superior.

- The outcome of eligible nonparticipants in ARDSnet studies is similar to the lower tidal volume group; therefore, ARDSnet did not prove that 6 ml/kg was any better than customized care. ARDSnet investigators responded that eligible nonparticipants are different from enrolled participants and the outcome is similar to both high and low tidal volume groups when an adjusted analysis is performed.

A number of very smart people are coming to very different conclusions analyzing these data. In retrospect, it is difficult to look for customization. We have to keep that in mind when we think about using unrestricted usual care as a control group – it will be very difficult to determine how clinicians were making decisions and thus to interpret such trials.

Three-Arm Trials and Usual Care in Effectiveness Trials

In a thoughtful piece in *Critical Care Medicine* last year, Drs. Silverman and Miller suggested that perhaps a three-arm trial, comparing both tidal volume strategies with a representative standard control group, would have offered the most clinical value by providing rigorous evidence to guide what should be considered the standard of care. Is this a win-win? Comparison of usual care to either tidal volume groups might answer the efficacy and effectiveness questions in a single trial and could protect research participants from anticipated harm should either of the two experimental arms demonstrate inferiority to usual care. This design might be reasonable if new therapies are not well represented in usual care and have risk. However, this design has some problems, one of which is the counterintuitive finding that such designs may be less safe under most assumptions.

If safety is defined as the additional deaths in the inferior treatment arm at the time it can be determined that this arm is inferior (i.e., when the study stops), three-arm trials with a usual care control usually result in more “additional deaths” (they are less safe) compared to two-arm trials due to the increase in sample size. Sample size is increased both because of the addition of the third arm but also because power is reduced from multiple comparisons during sequential interim monitoring, thus driving up the sample size further to maintain power. Safety is only improved if usual care is clearly superior to *both* of the experimental arms.

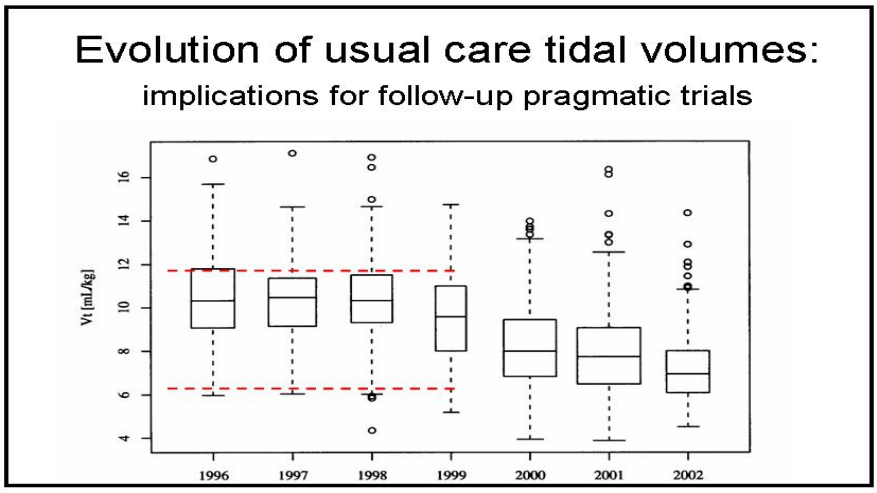
Usual Care in Effectiveness Trials

The design should fit the purpose. Usual care *should be* the control if the research question is to compare an intervention to usual care – the so-called pragmatic or effectiveness trial. An example in critical care is the process of knowledge development regarding weaning from mechanical ventilation. In the 1980s and early 1990s, clinicians were using many new weaning strategies made possible by new ventilator designs and better understanding of patient-ventilator interactions. Randomized efficacy trials compared four prevailing methods and identified the detection of weaning readiness (with a spontaneous breathing trial [SBT]) as the superior method. None of these studies controlled for customized usual care. These studies randomized to method 1 or method 2 or method 3, but not physician decision for best method. Advocates for usual care might have claimed they were flawed.

Effectiveness trials followed and there are now five randomized trials of protocolized SBTs versus customized, physician-directed care. Usual care in these trials was the desirable “control group” as this best fit the research question. The first of these effectiveness trials was published in 1996 and showed, as did most subsequent trials, that a nurse-initiated or therapist-initiated SBT protocol was superior to usual care practice. None showed superiority of usual care. These pragmatic trials helped the development of best practice guidelines now being adopted nationally.

It is interesting to note that these pragmatic trials do not age well. Usual care continues to change. Practice guidelines are moving usual care practices, which now incorporate

spontaneous breathing trials and other weaning methods (mostly checklists) into usual care in many ICUs, including the MICU at Johns Hopkins. A recent RCT at Hopkins now shows equivalence of protocolized SBTs to usual care.



How has usual care changed in ARDSnet hospitals? Are clinicians waiting for a followup effectiveness trial? Shown above are box plots of clinician-selected (usual care) tidal volumes prior to entry into ARDSnet trials in ml/kg predicted body weight from 1995 to 2002. The dotted horizontal red lines indicate the mean tidal volumes during the lower versus higher volume trial that finished in March 2000. It appears that most clinicians across these 18 U.S. centers and 1 center in Canada have decided to reduce tidal volumes. They have not waited for an effectiveness trial. Note, however, that had investigators planned a followup effectiveness trial of lower tidal volume to usual care, the usual care control group in 2000 would have been much different that the usual care control group in 2002. The ephemeral nature of usual care is its major limitation in clinical research.

Summary and Recommendations

Usual care should be the control group for effectiveness (pragmatic) trials; we would agree with Charles Weijer that this has primacy in that setting. When usual care is uniform (clinicians customize in similar and describable ways), then usual care ought to be the reference group. The problem in critical care is that substantial practice variation remains unexplained. In general, *de facto* usual care with substantial unexplained practice variability should not be the control arm in explanatory or efficacy trials. It is ephemeral, not reproducible, and difficult to pool across multiple trials. It is influenced by the unblinded research environment – many of these trials cannot be blinded – and variation introduces noise. There is no specific hypothesis tested *per se*, so if usual care is superior, it is unclear which of the usual care practices should be adopted. Also the converse of that is true: if usual care is inferior, clinicians can reject the new therapy

by arguing that their practice was not reflected in usual care. Ethical concerns arise when usual care practices are not judged to be reasonable, prudent, or competent.

We suggest that, in the absence of a well-defined clinical practice standard (wide variation that is largely unexplained), it is reasonable (Charles Weijer said “ethical”) to randomize patients to two well-founded yet competing beneficial treatment strategies that lie within the boundaries of competent or good care. Investigators, review committees, and IRBs must determine these boundaries. Obviously, equipoise and consent must also be present.

Three-arm trials with usual care as one of the arms are not routinely recommended. They tend to be larger and more expensive, and there is a high opportunity cost since most clinical trials are negative. Requiring the effectiveness comparison when efficacy has not yet been established is inefficient, would delay development of new therapies, and, under most circumstances, would result in more lives lost during the process of discovery.

How do you develop treatment strategies? They should be developed in an evidence-based fashion and independently reviewed by expert clinicians. As a general rule, Phase II testing for treatments outside usual care should be done. There should to be a consensus that both the treatments proposed in these explanatory trials are competent care.

How can risks in ICU trials of competing treatments be minimized? We are working in a practice environment that incorporates a wide number of practice styles. Many of these styles may not be best or prudent practice styles. We advocate for protocolized risk reduction. When possible, elements that reduce risks to participants in both treatment arms should be added. One example of this in the ARDSnet lower tidal volume trial was the inclusion of SBTs to both the higher and lower tidal volume groups. There were other scientific reasons for doing this – duration of mechanical ventilation was one of our endpoints and we needed to standardize weaning – but this also allowed us to offer a best practice to research participants.

The importance of the treating physician’s oversight is not much discussed or studied. The attending physician needs to be familiar with ICU trials in order to decide if either of the randomized treatments is prudent for the individual patient. It is an obligation of clinical trialists to ensure that intensivists practicing in the research environment understand the trial so they can make informed choices for their patients.

Such trials need sound DSMB oversight. One-third of those who agreed to participate in these trials (or the surrogates who agreed for them) will be dead in 28 days. These are high-mortality trials; it is important to be sure we have not strayed off the path. Near real-time outcome and safety data are recommended in these trials; for the most part, we use an electronic infrastructure for capturing these data. When the DSMB is thinking about whether randomized treatments are straying from usual care practices, a number of surrogate data can be examined, such as dropout rates, physician refusal rates, and physician modification or overrides of protocol instructions. These surrogates

can be monitored during the course of the trial as indicators of how well two treatments compare to contemporaneous usual care.

Selected References

The Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000; 342:1301-8.

Brochard L, Rauss A, Benito S, et al. Comparison of three methods of gradual withdrawal from ventilatory support during weaning from mechanical ventilation. *Am J Respir Crit Care Med* 1994; 150:896-903.

Carmichael LC, Dorinsky PM, Higgins SB, et al. Diagnosis and therapy of acute respiratory distress syndrome in adults: an international survey. *J Crit Care* 1996; 11:9-18.

Chastre J, Wolff M, Fagon JY, et al. Comparison of 8 vs 15 days of antibiotic therapy for ventilator-associated pneumonia in adults: a randomized trial. *JAMA* 2003; 290:2588-98.

Eichacker PQ, Gerstenberger EP, Banks SM, Cui X, Natanson C. Meta-analysis of acute lung injury and acute respiratory distress syndrome trials testing low tidal volumes. *Am J Respir Crit Care Med* 2002; 166:1510-4.

Esteban A, Frutos F, Tobin MJ, et al. A comparison of four methods of weaning patients from mechanical ventilation. Spanish Lung Failure Collaborative Group. *N Engl J Med* 1995; 332:345-50.

Gattinoni L, Brazzi L, Pelosi P, et al. A trial of goal-oriented hemodynamic therapy in critically ill patients. SvO2 Collaborative Group. *N Engl J Med* 1995; 333:1025-32.

Gattinoni L, Tognoni G, Pesenti A, et al. Effect of prone positioning on the survival of patients with acute respiratory failure. *N Engl J Med* 2001; 345:568-73.

Hebert PC, Wells G, Martin C, et al. A Canadian survey of transfusion practices in critically ill patients. Transfusion Requirements in Critical Care Investigators and the Canadian Critical Care Trials Group. *Crit Care Med* 1998; 26:482-7.

Hebert PC, Wells G, Blajchman MA, et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. Transfusion Requirements in Critical Care Investigators, Canadian Critical Care Trials Group. *N Engl J Med* 1999; 340:409-17.

Krishnan JA, Moore D, Robeson C, Rand CS, Fessler HE. A prospective, controlled trial of a protocol-based strategy to discontinue mechanical ventilation. *Am J Respir Crit Care Med* 2004; 169:673-8.

MacIntyre NR, Cook DJ, Ely EW, Jr., et al. Evidence-based guidelines for weaning and discontinuing ventilatory support: a collective task force facilitated by the American College of Chest Physicians, the American Association for Respiratory Care, and the American College of Critical Care Medicine. *Chest* 2001; 120:375S-95S.

Rogers EM. *Diffusion of Innovations*. 4th ed. New York, NY 1995.

The SAFE Study Investigators. A Comparison of Albumin and Saline for Fluid Resuscitation in the Intensive Care Unit. *N Engl J Med* 2004; 350:2247-2256.

Silverman HJ, Miller FG. Control group selection in critical care randomized controlled trials evaluating interventional strategies: An ethical assessment. *Crit Care Med* 2004; 32:852-7.

Van den Berghe G, Wouters P, Weekers F, et al. Intensive insulin therapy in the critically ill patients. *N Engl J Med* 2001; 345:1359-67.

Rivers E, Nguyen B, Havstad S, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med* 2001; 345:1368-77.

Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, Stern EJ, Hudson LD. Incidence and Outcomes of Acute Lung Injury. *N Engl J Med* 2005; 353:1685-16.

Thompson BT, Hayden D, Matthay MA, Brower R, Parsons PE. Clinicians' approaches to mechanical ventilation in acute lung injury and ARDS. *Chest* 2001; 120:1622-7.

Discussion

Q: Is it necessary for the researcher to be in equipoise about the care in both arms as opposed to his/her own care? I am thinking in the position of a patient. When I go to see one person and they recommend a tidal volume of 6 or 8, then I go to a colleague who is a "laggard" who says, "these things have changed but we do not know why, so I would recommend a higher tidal volume of 12 or 13." From my standpoint, if I cannot decide who is the more reliable physician and I have my experts recommending contradictory things, why would it be inadmissible for my treatment to be allocated at random rather than for me to randomly choose to believe A or B? Such a clinical trial would be run with the aim of resolving the conflict in the clinical community, but it may be the case that a number of those clinicians in it would have very determinate opinions about recommending one treatment over the other. What do you think?

Dr. Thompson: In essence, that recommendation will end up defining an equipoise that resembles the distribution of early adopters and laggards. The early adopters will not be comfortable with going back to the tidal volume of 12; they are done with that. The laggards will be uncomfortable with the new paradigms and will probably decline participation; they will not feel comfortable offering either of the two arms as treatment. If this operates in the ICU and early adoption and equipoise are linked, then the clinicians who participate in clinical trials are those in the early and late majority that still are not quite clear which treatment to offer. I do not know whether that is a valid construct, but we are asking for some reassurance that the application of these protocols is appropriate for the individual patient; that may or may not have been fully captured in the inclusion and exclusion criteria and in the design of the two protocols. There is an additional check – we make sure our overseeing physicians are educated about the nature of the experimental intervention.

Q: In its determination, the OHRP suggested a number of things. One idea was that the IRBs involved should have considered doing the survey of *de facto* actual practice in its hospital as part of its review of the protocol. In other words, they said there was potential value to the survey, but the survey would have been a survey looking only at the charts of that particular hospital, as opposed to all the hospitals in the network or just surveying people in general. What are your thoughts about that recommendation?
Dr. Thompson: Ultimately, OHRP found that the IRB process of reviewing this proposal was wanting and worked with us to provide a package of additional information that would allow the IRBs to reach a determination about where these studies sat in

proximity to usual care. We never answered the question as to whether it was necessary to document with chart review the exact mathematical description of usual care practices. In fact, OHRP relied on some qualitative expert judgment about usual care practices in the institution in which these studies were being tested. It was a decision about how robust the data are concerning what clinicians are doing in the environment.

At that level, just how robust do they need to be? This is an important question. Looking at the range of tidal volumes would be a first level approximation and you might say that is a reasonable request. But how do you get a customization? You have to do a recursive survey in which you move certain elements in and out of scenarios to find out what clinicians are conditioning on. It certainly is difficult to infer with scatter plots, as we just showed; it is a bit of a slippery slope. Fundamentally it is up to the IRB, and with a lot of latitude here, up to the judgment of expert clinicians in the community. The IRB found that the expert judgment sufficed in this setting, plus we provided fairly compelling evidence that usual care practices in general (not at the particular hospital) were quite varied.

This raises the question of local variations in care. Imagine you are looking at beta-blocker use for acute myocardial infarction. We saw this morning the map of that practice. If you are going to practice this in Texas, you have about a 50 percent chance of getting a beta-blocker after an acute MI. However, if you go to Dartmouth you have about a 97 percent chance. So if there is a usual care trial, IRBs in Texas and at Dartmouth have different issues. In that regard, it may be very important to know, particularly in a usual care trial, just what “usual care” is.

Case Presentation:
Case Study #1 – International Collaborative Ovarian Neoplasm (ICON) Trials of Ovarian Cancer Treatment

Ann Marie Swart, M.R.C.P., M.Sc.
United Kingdom Medical Research Council

Commentary:

Joseph L. Pater, M.D., M.Sc.
Royal College of Physicians and Surgeons of Canada

Benjamin Djulbegovic, M.D., Ph.D.
H. Lee Moffitt Cancer Center & Research Institute
University of South Florida

Case Presentation: Case Study #1 – International Collaborative Ovarian Neoplasm Trials of Ovarian Cancer Treatment

Ann Marie Swart, M.R.C.P., M.Sc., United Kingdom Medical Research Council (MRC)

Dr. Swart is a senior medical epidemiologist in the cancer division of the Medical Research Council's Clinical Trials Unit in London, United Kingdom. Dr. Swart is responsible for MRC's program of clinical trials in gynecological cancer and is a member of the U.K. National Cancer Research Network's Gynecological Cancer Clinical Studies Group and the International Gynecological Cancer Intergroup.

Epithelial ovarian cancer is the fourth most common cancer in women, but is not as common as bowel cancer or breast cancer. Most patients present with advanced disease; the survival rate is not as good as breast or bowel cancers but not as bad as some others. Twenty-five percent of patients present with earlier ovarian cancer, for which the 5-year survival rates are better (approximately 82 percent). Having said that, approximately 25 percent of patients relapse, and those who relapse usually die from their disease. These patients are certainly not all cured. Although it is the fourth most common cancer, it is still not that common; in the United Kingdom we have about 6,000 cases per year. It is a challenge for a single country (other than the United States) to randomize sufficient numbers of patients into clinical trials to get endpoints in appropriate timeframes.

The International Collaboration on Ovarian Neoplasms (ICON) originally was an informal collaboration involving the United Kingdom, Italy, Switzerland, and Norway. As clinical trials have become more formal, we have more formal collaborations with international groups. One of these groups is the lead group and is responsible for developing the protocol, the forms, and the database, and other countries join as participating groups.

Advantages of international collaboration for clinical trials include quicker recruitment and being forced to obtain an international consensus on practices within the trial and the trial objectives. One advantage is that either a positive or negative trial result is more likely to influence practice. Another advantage is that such an arrangement gives smaller countries, which would not necessarily be able to launch their own large-scale trials, the ability to participate in large-scale trials to improve important outcomes and survival of patients with ovarian cancer. Yet another advantage, especially in some of the latest studies including some of the molecular targeted agents, is that this collaboration has the potential to produce more leverage with pharmaceutical companies for access to drugs.

Many challenges of international collaboration include regulatory, financial, legal, authorship, and political issues. If you are trying to develop a protocol that will be acceptable to other collaborative groups, you have to make sure that all of the protocol-defined procedures and followup must be possible in different healthcare systems; otherwise, a group will not join your trial. It is not merely a choice of reference arms; there are all sorts of differences in medical practice among different countries. It is often the temptation of the individual groups to insist on their rigorous standards being

adopted by every single other group, but that is not practical. The approach we have taken for the important things is to define minimum acceptable standards and then allow flexibility within each country to include normal practice for them.

The first waves of the ICON trials are now closed; I will discuss those that are still relevant in terms of the reference arms they are informing. ICON1 and ICON2 were the first two trials; ICON1 was for early ovarian cancer and ICON2 was for more advanced disease. ICON3 and ICON4 were started when the Taxane question became of interest.

What drove the development of ICON1 and ICON2 was the first individual patient data meta-analysis, which formed in 1991. They got together and reviewed 45 trials, including 8,000 patients and 6,000 deaths and an additional larger meta-analysis. Until then, it was difficult to know what was standard in the treatment of ovarian cancer. These two meta-analyses answered the question – platinum was better than non-platinum for treatment of ovarian cancer. Platinum in combination was better than single-agent platinum, but at the same dose, cisplatin (CAP) and carboplatin were equally effective. The combination of three agents – cyclophosphamide, doxorubicin, and CAP – is better than cyclophosphamide and CAP. (This information will become important when discussing ICON3.) The meta-analysis also identified a real need for data on whether immediate drug treatment of early stage ovarian cancer was indicated; previously all those patients had been treated only with surgery.

I was not around when ICON1 and ICON2 were designed, so I can say that they were very elegantly designed! They included patients with ovarian cancer following surgery. So, following surgery, if the patient was found to have early ovarian cancer and the clinician was unsure whether to offer the treatment, patients were randomized to either immediate platinum-based chemotherapy with a specified minimum dose of platinum, or treatment that was delayed until clinically indicated.

ICON2 was developed for the remainder of the patients who had advanced disease and the clinicians believed that chemotherapy was required. These patients were randomized to single-agent carboplatin or CAP because CAP had been shown to be better than cyclophosphamide and CAP; however, it was unclear whether a higher dose of carboplatin given singly would be as effective as the combination. The problem with developing combinations is that, when new drugs come along, it is difficult to add another toxic drug into this complicated regimen.

ICON1 had to be combined with data from the EORTC ACTION trial. When we talk about international collaboration and getting patients in quickly, this is not a good example because it took 10 years to recruit about 900 participants. When the trial results were published, we were fortunate that the question was still relevant! We had a 33 percent reduction in the risk of death (which was highly statistically significant) and an 8 percent improvement in 5-year survival.

Since most patients received treatment with single-agent carboplatin, that should be standard treatment. If there were ever to be another early-stage ovarian cancer trial, we

would argue that that should be the reference arm. Early treatment trials for ovarian cancer are not planned at present.

ICON2 was the advanced disease trial, comparing single-agent carboplatin with the three-drug combination. This trial was stopped because the more important question had been whether taxanes should be introduced. The results of this trial confirmed that single-agent carboplatin given in optimum doses was as effective as the three-drug regimen.

In the mid-1990s, the taxanes became available, which led to the development of ICON3 and ICON4. Both ICON3, for advanced disease, and ICON4, for relapsed disease, have enrolled patients who require chemotherapy. These patients could be randomized to the reference arm or to paclitaxel and carboplatin. Because the results of ICON2 were not yet mature, it was decided that clinicians in ICON3 could specify before randomization which reference regimen the patients would receive. The reference regimens were carboplatin or CAP – the two arms of ICON2. These patients accrued between February 1995 and October 1998 with involvement from MRC, but we also randomized patients from Italy, Israel, Greece, and one other country. Originally, we understood that CAP was much more popular in Italy but only 25 percent of the Italian patients got CAP as opposed to one-third of the MRC patients. (I would be intrigued to look at this data over time to see if the ICON2 results changed Italian practice more than they did MRC practice.) We accrued a total of 2,074 patients in ICON3. Results became available in late 1998, and showed no difference in 2-year survival between the combined control arms and paclitaxel+carboplatin.

ICON3 was a large trial with no evidence of survival benefit and, looking at subgroups, no evidence that there were different effect sizes. These results were disappointing because we started this trial with the aim of improving outcomes for ovarian cancer. This led to years of debate about the interpretation of these data.

Looking at the paclitaxel controversy, we have ICON3 (a negative trial for Taxol), GOG132 (another negative trial for Taxol), and two trials in which there is a significant and impressive benefit. There are differences in the trials with respect to the reference arms. A naïve meta-analysis favors carboplatin+paclitaxel, but there is a lot of heterogeneity, thus questioning whether these trials should be put together at all. Using a random effects model, there is a benefit but it is not statistically significant. It is more important to explore the explanations for differences in the trial results, from the four trials, which was done neatly by J.C. Sandercock, who rationalized these four differences in the trials:

- Differences in extent of crossover (number of patients who had taxanes prior to progression or after progression)
- Differences in the types of patients (mainly in terms of residual disease after surgery)
- Differences in the treatment arms (3-hour or 24-hour paclitaxel)
- Differences in the control arms

Dr. Sandercock concluded that the only meaningful difference was that the positive trials all had the possibly inferior control arm of cyclophosphamide and CAP, which had come out of the meta-analysis conducted in the early 1990s.

ICON5 (in collaboration with the Gynecologic Oncology Group [GOG]), ICON6, and ICON7 are yet to be finished and/or published. ICON5 was developed as a joint study with GOG and MRC. We had our hands full at the end of the 1990s when the trial's development was going on, so we were probably less involved in the choice of reference arm than we should have been. The United States started this trial in January 2001 and we were not able to start it until the middle of 2002; our Italian colleagues started in the middle of 2003. So we had limited influence in developing ICON5, but I will go through how we rationalized our participation given that the reference arm in this trial is one that we did not feel was necessarily required.

The ICON5/GOG182 study design looks complicated but we were faced at the end of the 1990s with four interesting experimental regimens. We know that each trial takes between 5 and 10 years to conduct, and there would not be adequate time or resources to set up four individual studies. Therefore it was decided that there would be a multi-arm, multi-stage design that would incorporate all of the test regimens that looked interesting, each of which would be compared against the reference arm. We were criticized in the grant application because it looked so complicated, but that has not been a problem in carrying out the trial, as you will see from the recruitment figures in the United States. Normal outcome measures included overall survival and secondary outcomes. The statistical considerations supporting the idea that the multi-arm design was feasible were that the progression-free survival would be the primary endpoint for the first stage. Randomization would not be continued to any of the arms that did not have a hazards ratio of 1.15 compared to the reference arm. The idea was that you then have one or two arms to which you could expand recruitment. The final analysis of the main outcome was still overall survival and we were looking for a 9 percent increase in 3-year overall survival.

We believed that the ICON5 reference arm could have been carboplatin alone or carboplatin+paclitaxel because they had been shown to be roughly similar in ICON3. The way in which the ICON3 results were being applied with the U.K. clinicians is that there was a recommendation that the choice of treatment for all patients should be based on a discussion of the treatment options (carboplatin alone or carboplatin plus paclitaxel) with the individual patient. That is difficult to reconcile with a uniform reference arm for a trial.

Our survey of U.K. physicians had mixed responses. We floated the idea of a choice of carboplatin or carboplatin+Taxol prior to randomization. This conversation had to take place on a per-patient basis; it would have been better on a center basis but that was not feasible. We did discuss whether we could add a sixth arm of carboplatin alone; some people thought it was important to do this because otherwise we would be ignoring the results of our own study. In the end, we decided to stick with carboplatin and paclitaxel because the per-patient choice was not acceptable to our collaborators and the per-center choice was not acceptable to patients and clinicians. To add a sixth

arm at this stage was not going to work. The most important thing was to move on and test these new regimens.

More than 4,312 patients were recruited, with 3,435 from GOG. The interim analysis was performed last year and none of the test regimens jumped the hurdle—that is, to show a hazards ratio of 1.15. The independent data monitoring committee (IDMC) recommended that randomization need not continue. The results of overall survival will be mature next year and will be presented.

Where are we now? All the Gynecologic Cancer Intergroup (GCIG) groups got together. The proceedings of the 3rd International Ovarian Cancer Consensus Conference 2004 have just been published. We are trying to determine what is acceptable to standardize or what is acceptable to allow to be flexible in clinical trials of ovarian cancer. Recommendations are that chemotherapy should be standardized – the international standard comparator should be carboplatin+paclitaxel with a recommended minimum of dose of carboplatin and a flat dose of paclitaxel. Variations are allowed for clearly defined reasons; not every single ovarian cancer trial has carboplatin+paclitaxel in it. One study is looking at flat-dose carboplatin versus inter-patient dose escalation and many trials are still allowing choice, but for clearly defined reasons.

Large international trials are necessary in ovarian cancer. Despite variations in control and reference arms as well as research arms, we believe that it is still possible to design and run these trials in this context, with these choices, with minimum standards defined and with minimum specified doses. In Europe, the ICON trials have changed clinical practice, so we are satisfied with this approach.

Commentary on Case Study #1

Joseph L. Pater, M.D., M.Sc., Royal College of Physicians and Surgeons of Canada

Dr. Pater is a Fellow with the Royal College of Physicians and Surgeons of Canada. He has an M.D. and a Masters degree in Clinical Epidemiology and Biostatistics. He has been director of the National Cancer Institute of Canada and the Clinical Trials Group since 1980. He was professor and head of the Department of Community Health and Epidemiology at Queens University in Kingston, Ontario, and now holds the Edith Eisenhower Chair in Clinical Cancer Research at Queens.

The ICON trials are almost the only example in oncology in which choice of treatment regimen was allowed for the control arm. They raise some interesting issues; I will focus on the issues for design and interpretation of trials. I will tell the story from the North American perspective of what we thought we knew when these trials were designed.

ICON3 was designed to evaluate the role of a new drug, paclitaxel, in the management of ovarian cancer. It was designed at a time when we thought we knew some things, but we did not agree on what we knew. Despite the results of the meta-analysis,

however, there was quite a bit of uncertainty about what was the best non-paclitaxel containing regimen. Everyone agreed that cisplatin belonged in that regimen but not many people agreed with what else beyond cisplatin was necessary; the supporting evidence was debatable. In North America, the standard treatment arm was thought to be cyclophosphamide+cisplatin. At the same time, a trial that compared paclitaxel+cisplatin to cyclophosphamide+cisplatin (GOG111) had just shown a survival advantage for the paclitaxel-containing regimen, as did our trial (OV10). At the time that ICON3 was designed, the results of ICON2 (the comparison of carboplatin to CAP) were not available.

Two issues faced the individuals who designed these trials. First of all was what to do about the fact that results were not available from their own trial, which was evaluating either carboplatin or the combination of cyclophosphamide, doxorubicin, and cisplatin (the best non-paclitaxel regimen in the meta-analysis). The trial was designed to allow physician choice between these two regimens. As it turned out, the ICON2 results showed no difference between the two, so when ICON3 was analyzed, it did not matter that there had been a choice. It is unclear whether it would have been so easy to interpret ICON3, if ICON2 had turned out to show an advantage for either regimen.

The second issue was that the results of GOG111 and OV10 were considered irrelevant, that is, a non-paclitaxel containing regimen was used as a standard of care control arm. In oncology, positive trials are not common; two consecutive positive trials, both showing a statistically and clinically significant survival advantage, are difficult to ignore. Those of us who conducted those trials were taken aback that their results were ignored in the design of ICON3. As it turned out, we were wrong. For all the reasons stated, it turned out to be appropriate to do ICON3 the way it was conducted. But at the time, when the positive results of GOG111 and OV10 were becoming available, people thought we had established a new standard of care and wondered why these results were being ignored.

All of this led to ICON5. As was described, the goal was to move treatment forward. ICON5 aimed to compare a number of regimens incorporating novel agents to the current “standard of care.” The problem that ICON5 raised was that GOG111 and OV10 had shown a survival advantage of paclitaxel+cisplatin; ICON3 had shown that carboplatin was equivalent to paclitaxel+cisplatin. Unlike the situation with ICON3 in which we did not know the results of a key trial, in ICON5 we knew the results, but they were contradictory. A great deal of time on podium was spent trying to explain these conflicting results; everyone talked from the point of view of the trial they had conducted, which was positive, and came up with explanations of why other trials were negative. The most convincing explanation is the one you heard, but there is little evidence for dose responsiveness to cisplatin. There was considerable controversy about the explanation for these disparate results and about what was truly the best standard therapy. In general, North Americans favored paclitaxel-containing regimens whereas Europeans were less convinced.

The choice was to acknowledge the uncertainty and allow choice in the control arm or use the more widely accepted paclitaxel-containing regimen. In the end, the latter

choice was made and the trial was completed successfully. We do not know whether U.K. enrollment might have been greater if more choice had been allowed; the trial went so fast that it would have been difficult to catch up with the GOG in getting the study done. Depending on the results, interpreting them in the United Kingdom may be problematic.

The ICON trials are very appropriate for the purpose of this discussion. In oncology we have never really considered "usual care" in the way you have been talking about it today as a control arm. Having two control arms is a big stretch for us. ICON3 is one of the few examples of this type of design. The discussion as to whether the right choices have been made along the line in this series of trials will help us identify some of the issues we need to talk about.

Benjamin Djulbegovic, M.D., Ph.D., H. Lee Moffitt Cancer Center & Research Institute, University of South Florida

Dr. Djulbegovic is professor of medicine and oncology at the H. Lee Moffett Cancer Center and Research Institute at the University of South Florida in Tampa. His major academic research interests are in the areas of evidence-based medicine, decision analysis, clinical reasoning, systematic reviews, and meta-analysis and the ethics of clinical trials. Most recently, he has focused on the understanding of the relationship between ethical precepts and the innovative successes in clinical research.

In my presentation, I will focus on uncertainty, how we acknowledge clinical uncertainties, and how we address those uncertainties. My comments will revolve around the need to articulate uncertainties and to show that acknowledgment of uncertainties around competing treatment alternatives translates into the mechanism for choice of the comparator intervention. My aim is to provide an explicit framework to the process that appears to have been followed implicitly in the ICON5/GOG-182 trial.

According to current research, the most important feature in the design of clinical trials may be the choice of a control intervention, likely overriding all other important aspects of trial design. The study can adhere to all contemporary recommendations for good design and still produce predictably biased results.

Is there a formal mechanism to select an appropriate comparator similar to using randomization to control selection bias? This mechanism revolves around equipoise or the uncertainty principle, which states that a clinical trial is ethically or scientifically justified only if there is substantial uncertainty concerning which treatment is more likely to benefit patients. Ultimately, this principle relates to the choice of a control intervention.

Whether there is uncertainty or not depends on the state of accumulated evidence. It should be obvious that if our pre-existing knowledge determines one treatment to be superior over the other, then research would not be justified on either ethical or scientific grounds. This has been recognized for some time. However, what is needed is a blueprint for action in order to further operationalize some of these general issues.

The first operational issue relates to the question of which method is best when addressing uncertainties with regard to relative effects of competing treatment interventions. There is an increasing consensus that the use of methodology of systematic review and meta-analysis should be employed as the first method of choice in addressing uncertainties regarding effects of therapeutic interventions. These methods provide research synthesis of totality of evidence on the treatment effects, and in turn can inform development of practice guidelines. The second method in addressing the state of existing uncertainties can be based on formal surveys or audits of expert clinical practitioners. Finally, somewhat less discussed but equally important is publication of a trial protocol to solicit critical appraisal, which may give us further insights related to the contemporary views of the best existing treatment alternatives.

Let me first say something about the use of practice guidelines to define the standard of care. Ultimately, practice guidelines represent the final link between basic and clinical research in our pursuit to improve patients' health outcomes. In clinical practice guidelines, only successful, effective interventions with benefits outweighing harms would be recommended. Therefore, it is possible to use guidelines to determine the standard of care. For example, in 2005, the National Cancer Center Network (NCCN), an organization composed of 20 leading U.S. cancer institutions, listed in their guidelines three equally acceptable alternatives as the drugs of choice for managing ovarian cancer.

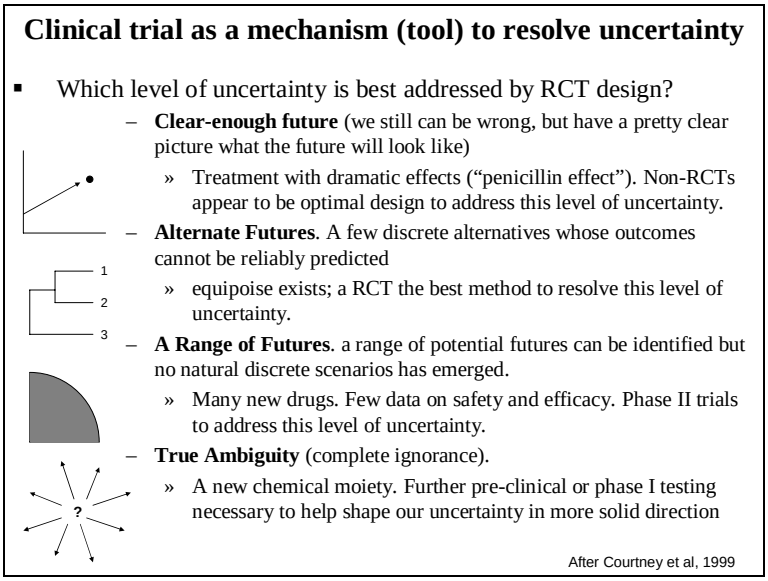
If all of these techniques (systematic review, practice guidelines, survey of practitioners, and publication of research protocol) show that one comparator is clearly favored or there is substantial uncertainly or disagreement among expert clinicians concerning the merits of proposed interventions, it is likely that an inferior comparator would not be chosen for the trial and therefore the proper choice of a control intervention can be said to be affirmed. These considerations can be succinctly expressed in the form of the following question: "Were the investigators uncertain [about the choice of control intervention] or were they in equipoise when they designed the trial?"

In fact, uncertainty of practitioners was used as an eligibility criterion in the ICON3 trial.

When talking about uncertainty, it needs to be said that uncertainty comes in different shapes and forms. It can range from simply lacking the factual confirmation of an otherwise clear understanding of treatment effects to what we like to call the maximum uncertainty or equipoise. The crucial issue is how we choose a scientific methodology to address the underlying level of uncertainty. Of critical importance in advancing this debate is the development of the taxonomy of uncertainties in clinical medicine. The purpose of science in general, and clinical trial methods in particular, is to address (resolve) the existing uncertainties regarding the effects of competing treatments. The underlying idea is that the choice of scientific methodology and analytic technique should be tailored to the underlying uncertainty about treatment effects. One (methodological) size does not fit all (levels of uncertainties).

Since no taxonomy of clinical uncertainties has been developed to date, for the purposes of clinical use I adopted (see Fig 1) the classification of scientific uncertainties from economic literature.

Figure 1. Taxonomy of clinical uncertainties
The figure illustrates how choice of scientific methodology is tailored to underlying uncertainties about treatment effects.



When formulated this way, the question now raised is, "Which level of uncertainty is best addressed by a comparative randomized controlled trial (RCT) design?" Using this taxonomy of uncertainty, let me walk you through figure 1, focusing on the clinical trial as a tool to resolve uncertainties about treatment effects:

- "Clear-enough future" scenario – We still can be wrong but we have a pretty clear understanding of what the future will look like. This scenario would apply to those occasional situations in medicine in which treatments are associated with dramatic effects. Examples of such situations are penicillin for pneumococcal pneumonia, insulin for diabetic coma, and red blood transfusion in bleeding patients. This is a clinical situation when a noncomparative trial will suffice. We do not need an RCT to address this level of uncertainty or to provide additional evidentiary support for using these treatments under these clinical circumstances.

- "Alternate futures" scenario – There are few discrete alternatives that have emerged but whose outcome we cannot reliably predict. That is when the equipoise requirement is met and an RCT would be the best method to resolve this level of clinical uncertainty. This talk focuses on this level of clinical uncertainty.

- "A range of futures" scenario – There is a range of potential futures that can be identified but no natural discrete scenario has yet emerged. For example, currently in oncology there are many new drugs but very little data on their safety and efficacy. Typically, we employ Phase II trials to address this level of uncertainty.

- "True ambiguity (complete ignorance)" scenario – This is often seen with a new chemical moiety and we have no idea how any of these drugs will work in the clinic, so more data is needed in preclinical and Phase I testing to shape uncertainty in a more solid direction.

We have conducted a lot of research in the last several years which can be summed up as: an RCT should be considered as the method of choice to resolve uncertainties when alternatives are clearly formulated and when there is about equal chance (in terms of benefits and risks) that one treatment is better than the others. As stated above, this can be dubbed as equipoise. It is also important to remember that uncertainties are not fixed in time; uncertainties may shift and migrate over time. Typically in clinical medicine, this happens when the scenario "range of futures" solidifies into the "alternative futures" scenario, with two or more treatment alternatives culminating in equipoise, at which point we would use an RCT as a method of choice to address this newly articulated uncertainty.

When applying these concepts to the GOG-182/ICON-5 trials, and as discussed by Dr. Pater in the 1990s, two identified standards – carboplatin+paclitaxel and single-agent carboplatin – clearly emerged as acceptable new standards. However, these new treatments, as good as they were, have not been successful in curing ovarian cancer. Outcomes remain inferior and the median survival of patients with advanced ovarian cancer is only about 36 months. Therefore, the need for new therapeutic agents emerged. New agents just appearing on the clinical scene when the design of the ICON5-GOG182 trial was formulated were gemcitabine, pegylated liposomal doxorubicin, and topotecan, all of which showed activities as a single agent in Phase II studies. None of these agents demonstrated platinum cross-resistance, which was deemed to be very important as a biological rationale for use in design of the ICON-5/GOG-182 trial. From the point of view of the choice of an adequate comparator, the key point here is that there was no doubt, i.e., there was no uncertainty that as single agents these drugs were expected to be superior to standard treatment. However, if combined with the standard treatment, the investigators believed that these new drugs might offer some additional benefit.

This is what I think the ICON-5/GOG-182 investigators faced when they designed the trial and chose the comparator interventions. Expressing it in the language shown in Fig 1, they had clearly established treatment standards ("alternate futures") that they needed to reconcile with emerging new drugs ("range of futures") in the design of the ICON-5/GOG-182 trial. In the 1990s, platinum-based therapy plus taxanes or carboplatin as a single agent were established as treatment standards. The FDA approved cisplatin and paclitaxel as the first line of therapy. The GOG showed that cisplatin and paclitaxel were about equal to carboplatin and paclitaxel. We also heard

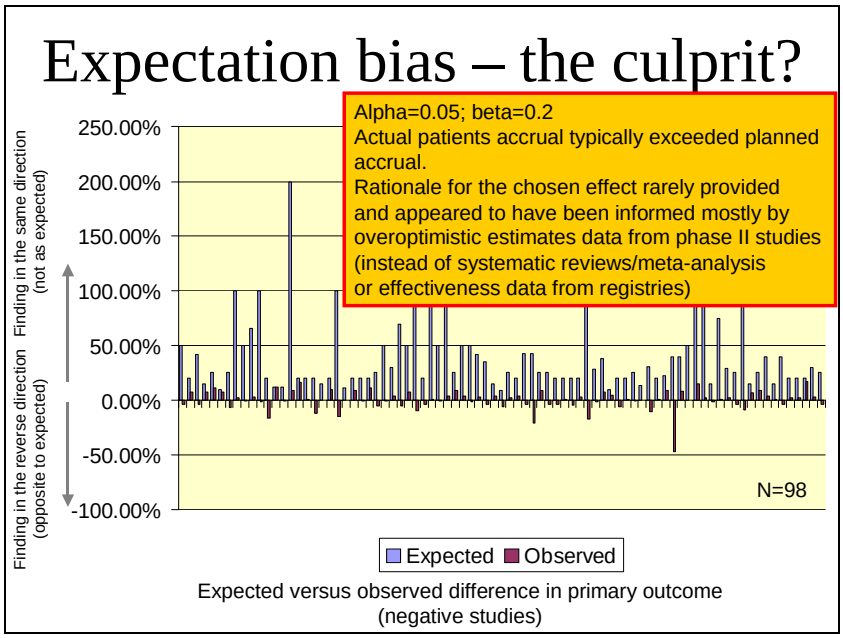
from Dr. Swart that carboplatin+paclitaxel or carboplatin alone could also have been considered as established standards, based on the ICON-3 trial meta-analysis showing that carboplatin and paclitaxel are equally acceptable treatment alternatives and a survey of the U.K. practitioners. Finally, an international conference consensus affirmed carboplatin and paclitaxel as a standard of care.

Against this background, one should understand emerging data with new agents ("range of futures"); gemcitabine, doxorubicin, and topotecan were all showing promising activity in early Phase II trials. Gemcitabine was tested in three Phase II trials with carboplatin and paclitaxel and was well tolerated and considered effective; therefore, the rationale was made that it could be combined with carboplatin and paclitaxel. Doxorubicin was effective in ovarian cancer and was used with carboplatin and paclitaxel as the experimental arm in one of the GOG trials; therefore, that rationale was provided to use this regimen as one of the competing treatment arms. Topotecan was approved by the FDA as second-line therapy, but it was found to be too toxic when given simultaneously with other chemotherapy agents. Some data existed that indicated that, if given sequentially, topotecan might be more effective and less toxic; therefore, two other treatment alternatives were formulated to be used in the ICON5/GOG0182 trial.

As you can see from "the range of futures" to "clearly established alternatives," five equally uncertain alternatives were formulated, each of which was equally believed to be the superior therapy ("the winner"). Hence, a rationale for ICON-5/GOG-182 trial was made. That is, the choice of comparators was affirmed by systematic review of existing evidence, practice guidelines (including the FDA approval of one of the treatment arms), surveys of practitioners, and peer review of the research protocol. As a result, the ICON5/GOG-182 randomized trial was considered ethically and scientifically justified to resolve the uncertainties with respect to which one of these five treatment alternatives is superior.

Thus far, I have been discussing the choice of a comparator in a qualitative sense. Ultimately, the design of the trial must integrate a null hypothesis in quantitative terms, specifying the differences between treatment effects that one would like to detect. What was uncertainty about the effect size in the ICON5-GOG-182 trial? What was the rationale for the chosen effect size? How was the choice informed? We heard that in the ICON5-GOG-182 trial, it appeared that the choice had been informed by the totality of evidence accumulated during prior ICON trials, which provided the basis for sample size calculation. However, in many trials this is often not the case.

At the Cochrane meeting, we recently presented our analysis of 150 trials performed under the auspices of the NCI, all of which were high-quality trials but which produced findings that were either inconclusive or negative. When we looked for the usual reasons why these trials generated inconclusive or negative results, such as poor accrual, no explanation could be found. It appears that the main reason for inconclusive results was that the investigators were overly optimistic regarding the treatment differences they were hoping to detect.



As you can see in the above slide, what they expected and what they observed never matched. Often, what they observed went in the opposite direction of the investigators' expectations. A rationale for chosen effect size was rarely provided and, when provided, it appeared to have been informed mostly by exceedingly optimistic estimates from Phase II trials instead of systematic reviews/meta-analyses of the totality of research evidence accumulated prior to initiation of the particular trial.

In my final considerations, I want to focus on whether one can expect the ICON5-GOG182 trial to address all uncertainties around all possible treatment alternatives that exist in the management of ovarian cancer. Of course not. Here, I would like to draw your attention to different research interests in health care that exist among its multiple constituencies. For example, if I am a patient and my doctor says he will give me treatment A and then I go across the street and another doctor says he will give me treatment B, I am then interested in having the answer to this uncertainty: which of these treatments, which have been used in the practice of medicine for a long time but never directly compared, is better? But if I am a researcher, I will not likely be interested in designing the study that will test the effects of two alternatives that have been around for 10 or 20 years; I would more likely be curious about new, more exciting treatments. Therefore, I would probably try to design the trial to help resolve uncertainty about new treatments versus old. If I were a sponsor I would be more inclined to choose a comparator that would offer me the highest likelihood that my treatment will turn out to be "a winner" and hopefully get approved by the FDA. If I am a policymaker, I would like to compare those treatments that would help me make decisions regarding the most cost-effective alternative for the majority of patients.

So the question about the choice of comparator intervention can be formulated as a function of multiple interests in the setting of the research agenda. Indeed, the question can be asked, "Whose uncertainty counts?" Who is setting a research agenda – patients, researchers, regulators, sponsors (pharmaceutical industry)? And to borrow

from Kenneth Arrow, a Nobel Laureate in economics, who, in a different context, showed a long time ago that when there are multiple choices of interests, it is impossible to provide consistent, unambiguous decisions that will be equally acceptable to all parties – in our case, these various constituencies consist of trial patients, future patients, researchers, policymakers, and regulatory agencies.

The best we can do is to acknowledge our uncertainties and provide transparent and explicit evidence-based review of our existing knowledge. I would argue that when this document comes out it should state explicitly that we need to articulate uncertainties regarding chosen comparators and endorse the methodology of systematic review of the literature to synthesize the totality of research evidence in an attempt to account for what is already known. Ultimately, the issue concerning "usual or standard of care" relates to the question of assessment of the existing knowledge. Methods of systematic review, along with other methods discussed here, is the best we can do to ensure the choice of an appropriate comparator. No trial should be approved before a systematic review of the literature is conducted. (My understanding is that this requirement already exists in several European countries, although I am not sure how well it is enforced).

We should recognize that no magic formula is possible in selection of the comparator intervention(s) that will be equally acceptable to all parties. The best we can do is to provide transparency in decisionmaking, clarity regarding the criteria that we are putting forward (as illustrated in the ICON5-GOG182 trials), shared deliberation, and the chance for an appeal process (via publicly available research protocols). The issue is as equally ethical as scientific. This process is analogous to other choices that involve multiple constituencies, also known as "accountability for reasonableness" after Daniels and Sabin. The approach could help us legitimize specific choices that may favor one set of stakeholders over others. The tension due to choices that have to be made cannot be completely eliminated but, if the choices are deliberated in the way just suggested, our explanation will be better understood and in the end better accepted by various parties interested in the design of the trial.

Panel Discussion of Case Study #1

Dr. Wittes: I am impressed with the series of ICON trials and want to comment from the vantage point of someone to whom this is new.

One of the issues that this series raises is the particular problems associated with a limited horizon of patients. It is not a rare disease, but it is not a disease that has huge numbers; therefore, the necessity to think about efficiencies of design is crucial and laudable. The ICON3 paired control groups is what I called in my presentation "a menu of options"; it was a short menu of two, with both controls clearly defined, and it was done in the "spirit of the best" as Helsinki would say. The design was chosen by selecting the two treatment options that, at the time, were viewed as the best current care. As it turned out, they were indistinguishable in terms of efficacy.
The other nice part about the development of these studies is the sophistication and thoughtfulness of putting together the trial, in a rational and thoughtful way –

investigating the sources of commonalities and the variables that seemed to make them different, and the attempt to try to tease that out. I am particularly sensitive to that because some of us are trying to do that in Celecoxib studies, and it is very difficult. When you have a small number of trials with slightly different doses and other variables and largely different answers, it is difficult to figure out what is going on.

I also want to raise the issue of crossover. Dr. Swart mentioned it a little bit in her talk, it was more explicit in the material related to the ICON trials that she sent to us, and it was brought up earlier today. Statisticians are likely less concerned about crossovers than much of the rest of the world, but there are three positions that people take. This is the question of people—in the study testing whether Paclitaxel conferred some advantages over the control arm—crossing over to Paclitaxel, and the question, therefore, as to whether that ruined the study. Some people take the position that crossover makes studies "no good" and you have to wash away the results that you cannot interpret. Others take the point of view that, if there is crossover, you censor analysis at the point of crossover and the informative information is what you gather up to the point of crossover. Others of us take the more radical position that crossover can be viewed, in many cases, as part of the trajectory of the disease and the treatment. If you are randomized in the beginning, that is when the groups are the same; then they diverge, and part of diversion is taking on different treatments. It is a plea not to throw up your hands every time crossover occurs.

The estimate for ICON5, looking for a 33 percent reduction, seemed a little optimistic; the only study that had shown such a reduction was ICON1. By analogy it seemed big, but maybe one does not want to start the kind of chemotherapy one is talking about without such a big reduction. In general, having the guts to do a five-arm trial, which on first blush looks complicated but then it is simply adding or subtracting one agent or another, has the potential to answer many important questions.

From your talk, I did not understand why a per-patient choice in ICON5 would not have been acceptable statistically.

Dr. Swart: It was not acceptable to GOG.

Dr. Levine: Given the profound differences between the United States and the United Kingdom in what constitutes usual care, does it make sense to have an international collaboration in which, almost certainly, some of the people will not be satisfied or will not find the result useful in own practices?

Dr. Swart: It is probably true of the earlier trials, because we were perhaps more pragmatic in terms of not worrying about stage of disease. If a patient required chemotherapy they could go to ICON2, and if it was not clear whether they needed chemotherapy they could randomize to ICON1. We moved away slightly from that level of pragmatism. There is no more standardization of staging of minimum standards of surgery. The beauty of the intergroup is that we have a lot of opportunity to decide on what bothers us all most and to agree on minimum standards. In theory, everyone should be happy.

Dr. Pater: At NCI Canada we do a lot of international studies with Europe and with the United States, and although there are variations, those variations may not be as great between countries as it appears regarding what people think is the best current therapy. There is a lot of variation in standards of surgery and other things. Most of us designing large relatively pragmatic trials agree that the comparison is on that background of variation and, as long as it is a randomized trial, we can draw reasonable conclusions. The ICON1 action results might be an exception to that. Ann Marie presented them very clearly and I think this is the proper interpretation, but there are a lot of people who point out that the patients in the United Kingdom in that study did not have lymph node dissection and laparotomy, and that the benefit of chemotherapy seems to be greater in patients who did not have as extensive surgery. That is an unfair subgroup analysis but it has had an impact on interpretation of the results. Overall, many trials are conducted internationally and are persuasive in most places.

Dr. Wittes: In ICON5 you talked about how recruitment was so quick. Have you been able to maintain the five different regimens?

Dr. Swart: On the interim analysis, none of the experimental arms reached across the hurdle and therefore recruitment has stopped. Yes, people actually adhere to the regimen. Part of the informed consent process was a general patient information sheet and then, depending on the arm the person was randomized to, there was additional information. They could all see what all five regimens looked like beforehand, if they wanted to. Patient recruitment did not seem to be an issue. We are now doing a similar prostate study, with Canada.

Q: Is it safe to assume that there were no important differences in adverse effects between any of these arms?

Dr. Swart: The DMC recommendation not to continue was because of no demonstrated efficacy; there was no comment on the safety profile but that will come out. They had to jump the efficacy hurdle as well as have an acceptable safety profile.

Q: The field of oncology trials might be a good place to bring up the question of measuring and simultaneously considering outcomes in different dimensions. In oncology traditionally, it is fairly rigorous measures of survival or disease-free survival, but there are also differences that can be pronounced in side effects or quality of life. It was interesting to see that the Taxol-based and platinum-based results seem comparable in terms of survival, but my understanding is that there are significant differences in terms of side effects – renal toxicity, rescue from infections, etc. Had anyone thought of incorporating some of these other dimensions that involve patient preferences (quality of life or disability-adjusted life) into these trials, retrospectively? How much more of a burden would it be to add that kind of thing prospectively?

Dr. Swart: QOL and health outcome are part of all of these large-scale trials. When no clinical benefit is seen in survival, then the QOL assessment tends not to be so interesting.

Dr. Pater: There was a QOL component to the OV10 trial. It did not show appreciable differences in QOL between the Taxol and cyclophosphamide arms. That is not surprising, since Taxol and cyclophosphamide are [not] that much different; it is the platinum that is the problem because it is a component of everything.

Q: Regarding effect size and sample size, especially in surgical trials, I have had advice in two different directions. In one direction to look for large differences that will make people change clinical practice. If there is only a 10 percent difference, people will keep doing what they want to do anyway; a 20 percent or 30 percent difference will make people change what they do. The other advice is that you do not want to miss a difference if it exists. In determining the control group, how do you keep in mind the effect size and the sample size?

Dr. Djulbegovic: This will have to be debated because you are trading off false positives for false negatives – which one do you value most? Do you want to discover breakthroughs and then, in our assessment in our trials a large effect size seen between 7 percent and 15 percent of cases means the majority of other results will be false positive. You risk that you will recommend ineffective treatment to future patients. Compare that to whether you are going to miss some important effect, if you are talking about frequent disease on a large scale, that will save many lives. It really comes to the value of decisionmakers and policymakers in terms of trading false positives versus false negatives.

Dr. Pater: NCI/Canada, CCTG, and with cooperation from the United States, we have conducted many clinical trials and we always think the control arm is standard care, at least *an* accepted standard. What is happening in oncology trials is that we are now experiencing what happened in cardiology trials 20 years ago – we thought a large trial was a few hundred people. The rationale for that was that it would take a very large difference to adopt these horrible treatments. Now we are doing breast cancer trials in which we are trying to tease out the difference between different kinds and ways of inhibiting estrogen production on the survival of breast cancer patients, and now we are doing 6,000-patient studies. The principle is the same – we want to do a study that targets a plausible difference such that, if it were large enough, the therapy would be adopted. Oncologists are always thinking about changing practice when they do their studies.

Dr. Wittes: It depends on whether you are thinking about a new treatment that will change practice or whether you are thinking in terms of incremental changes. When you think about incremental changes, you think of realistic effects that will be small. One of the problems in oncology is that every incremental change represents thousands of dollars.

Q: Aside from being overwhelmed with a five-arm control group trial design, I wondered how you addressed two issues. One is, recognizing that people in different countries come from different cultural backgrounds and likely have practices that are different in many ways that may be difficult to articulate, did you perceive the different choices that physicians in those countries preferred regarding drug use as surrogates for different

strategies of care that included other elements? What did you do to identify those elements? If an Italian wants to use drug X and a German wants to use drug Y, are those surrogates for different attitudes toward care in general?

Secondly, why did you allow clinicians to choose their control drug rather than developing and making part of the protocol the rules to identify how they should allocate the drugs that were preferred in their environments?

Dr. Swart: We work by relying on collaborations with the Intergroups to judge whether a trial will be acceptable in their country. We work with other trialists, who have to understand the protocol, and we highlight any limitations or difficulties we may need to address. That is how we get around differences in cultures and attitudes and what people might want to use. Then perhaps they want to use more than six cycles so we have to discuss whether it is appropriate to allow more than six cycles of chemotherapy; if it is, we try to accommodate it and if it is not, then they can choose whether to take part or not. We always stratify by country or group to take account of differences.

Because we had ICON2 that was asking which of these regimens was better, that is how we justified the physician choice in the reference arm of ICON3. Although it did not occur for earlier trials, for ICON6 and ICON7 we have gotten together the various country groups to talk about how they manage and monitor patients independent of the drugs chosen.

Q: We have heard it said several times that the IRB is responsible for passing on the scientific merit or design of these protocols. When you are faced with questions of this sort – for example, what to use as a comparator – do you turn to the IRBs and ask them for their input?

Dr. Pater: In the North American Cooperative Group setting, there is a central IRB. The way the process occurs now is that all new protocols sponsored by NCI must go through both the internal CTAP review and then the IRB review. So in a sense, that does happen and that can be a very iterative process.

Dr. Levine: You are dealing with a very special IRB that is designed to be especially competent in reviewing protocols in the field of oncology. At most of our institutions, the institution may have one or two experts in the particular field of study. Our rules say that, in case they happen to be the principal investigator, which is usually the case, we have to excuse them from the discussion. IRBs are by design incompetent to do scientific review.

Q: I will speak for the IRB perspective. We have cancer trials, we heard about the ARDSnet, and now the IRB has to look at what is competent care for X, Y, and Z diseases. Do not come to the IRB for this. The IRB is only one piece of a human subjects protection program. We rely heavily on the signoff of our department chiefs, who say whether this is a reasonable experiment to do in their setting. The IRB cannot even pronounce some of these words, let alone have an opinion on them, as well as

whether a tidal volume of 6 or 12 is going to be appropriate. We are one piece but not the final answer; it has to be all pieces together.

Q: In the five comparators and the new treatment, did you consider comparing the new treatment to the combined results of all of the five? If not, would the panel give some comments on the advantage or disadvantage of doing that comparison? Standard care is whatever it was in those different countries.

Dr. Swart: There is a reference regimen of carbotaxol and four experimental arms. There is no plan to combine those four experimental arms and compare them with the reference arm. They are very different regimens with different toxicities.

Q: You could also say that this is standard treatment in those different places and you could compare the new to that standard treatment.

Dr. Wittes: My understanding is that those four are the four new ones; the one is the control.

Case Presentation:
Case Study #2 – Multimodal Treatment Study of ADHD (MTA)

James Swanson, Ph.D.
MTA Cooperative Group, Yale University

Commentary:

Julie Magno Zito, Ph.D.
University of Maryland
Constantine Frangakis, Ph.D.
Johns Hopkins University
Paula D. Riggs, M.D.
University of Colorado Health Sciences Center
Betty Tai, Ph.D.
National Institute on Drug Abuse, NIH
Charles Weijer, M.D., Ph.D., FRCPC
University of Western Ontario

Case Presentation: Case Study #2 – Multimodal Treatment Study of ADHD (MTA)

James Swanson, Ph.D., MTA Cooperative Group, Yale University

Dr. Swanson is a developmental psychologist and professor of pediatrics and director of the UCI Child Development Center, which he founded in 1983. At that center, Dr. Swanson established a large clinical treatment program for school-aged children with attention deficit hyperactivity disorder (ADHD). He led the research team for one of the six sites of the National Institute of Mental Health-funded Multimodal Treatment Study of ADHD (MTA) and for one of the six sites of the Preschool ADHD Treatment Study (PATS).

The Multimodal Treatment study of ADHD (MTA) addressed treatments of a behavioral disorder most recently described in the fourth revision of the Diagnostic and Statistical Manual (DSM) of the American Psychiatric Association (APA), published about the same time (DSM-IV, 1994) the MTA study started. Attention Deficit Hyperactivity Disorder (ADHD) has two symptom domains – inattention and hyperactivity/impulsivity. Each domain consists of nine behaviors that may appear to be just characteristics of childhood (e.g., "has difficulty sustaining attention," "leaves seat in classroom," etc.) but are considered symptoms of a disorder when they occur at a high intensity and frequency that is much greater than other children, and as a result produce significant impairment in the child's life.

The primary treatment of ADHD and its predecessors (e.g., "hyperkinetic reaction of childhood" or "minimal brain dysfunction"), dating back more than a half century (Bradley, 1937), has been with stimulant medication. In 1990 the U.S. Department of Education funded a center at UCI to conduct a review of all articles published on the use of stimulant medication to treat ADHD. A rigorous review revealed a literature consisting of about 4,000 articles. Fortunately, there were then already 341 reviews of the literature, which allowed our center to write a "review of the reviews" summarized the empirical evidence of the efficacy of this treatment (Swanson et al., 1991). The evidence is very strong, from many clinical studies and controlled trials, that stimulant drugs have beneficial effects by reducing the symptoms of ADHD and some associated features often present such as oppositional, defiant, and aggressive behavior. However, the "review of reviews" also revealed that almost all studies were of short-term effects; very few studies were of long-term effects, and this was a significant concern. "Short-term" could mean as short as effects of a single dose over 3 or 4 hours, but most were of the effects of treatment for a few weeks or months. Another complicating factor was a changing pattern of clinical practice. Over many years, multiple publications have documented the increase in the clinical use of stimulant medication, which was detected initially in the 1970s (Krager and Safer, 1974) and has continued linearly to the present (Zito et al., 2003). So, over time the usual community treatment has been changing, due to an increasing use of stimulant medication to treat ADHD.

When the MTA study began, methylphenidate (Ritalin) was the primary stimulant medication used for the treatment of ADHD, although d-amphetamine (Dexedrine) was also used. The half-life of methylphenidate is short (about 2 or 3 hours), so in usual

community treatment settings it was administered two or three times a day. The sustained-release formulations of methylphenidate available at this time were not considered effective and therefore were not prescribed very often. This standard treatment regime required participation of school personnel to administer the medication at school, since methylphenidate and amphetamine are Schedule II drugs and requires tight control because of the potential for diversion and abuse.

In addition to stimulant medication, behavior modification had been used for psychosocial intervention in community settings of the school and home. One way to provide this intervention is through parent training, in which parents are taught the basic principles and techniques of behavior modification – reinforcement, extinction, punishment, and stimulus control. In the MTA, a very intensive parent training course focused on the application of behavior modification methods in the home setting. Early in the stages of the behavioral treatment package, an all-day 8-week summer treatment program started the children with an intensive experience. In the school setting, another direct intervention involved paraprofessionals (trained classroom aides) who worked with the teacher to apply behavior modification based on a token system in which children earned check marks or points for appropriate behavior. A daily report and home-based reward program was used in which children could earn check marks during each classroom period, and at the end of the day the total could be sent home and access to natural reinforcers were provided in conjunction with the parents' component of the behavior modification program.

The MTA was a cooperative agreement among multiple sites at universities and hospitals (UCI, Duke University, University of Pittsburgh, Columbia University, UC Berkeley, and Long Island Jewish and Montreal Children's Hospitals) and the program office at the National Institute of Mental Health (NIMH). This type of multi-site study is considered appropriate when there is a significant public health issue at stake, when single site approaches are insufficient, and when established treatments (in this case, medications and behavior modification) are widely used in the community but adequate evidence of relative effectiveness has not been established by controlled studies.

The MTA started with a planning and protocol development phase in 1993. Only children with DSM-IV ADHD-Combined Type were recruited for the study, which requires the presence of both domains of symptoms (inattention and hyperactivity/impulsivity). One of the major problems in the literature was the uniqueness of some samples due to variation in local conditions and referral biases. In one study children might be referred to a Department of Psychiatry in New York City, while in another to a Department of Pediatrics in Irvine California. The six-site MTA represented a cross-section of the types of settings where treatments of ADHD were provided. The sites also used newspaper and radio advertisements as well as usual referral process for identifying children for the study.

A randomized clinical trial design was adopted for the evaluation of two established unimodal treatments, medical management (MedMgt) and behavior modification (Beh), compared to each other as well as to multimodal treatment defined by the combination (Comb) of the two treatment modalities. In the first phase, MTA staff provided intensive

treatment for a 14-month period of time, which was considered to be relatively long-term treatment compared to previous studies. Algorithms for implementing the intensive treatments were developed by strong advocates in the study -- physicians who were strong advocates of medication (MedMgt) as the first line of treatment, and psychologists who were strong advocates of behavior modification (Beh) as the first line of treatment. The medication treatment selected for the MTA was three-times-a-day dosing with methylphenidate given seven days a week, with the option of trying other proven medications if methylphenidate was not satisfactory. The behavior modification treatment was based on multiple components that were integrated over two settings (home and school) and included parent training, the UCI paraprofessional program for 4 hours per day for 12 weeks, the daily report and home-based reward program implemented at the end of one school year and the beginning of the next, teacher consultation during these two school years, and a summer treatment program in a camp-like setting for 8 hours a day for 6 weeks between the two school years. This psychosocial intervention was more intensive than typical in usual community settings. Several of the investigators had developed components for this type of intensive intervention. By combining these some of us thought the intensive behavioral approach (Beh) would be equal to or even superior to treatment with stimulant medication (MedMgt) and some thought the reverse would hold. Almost everyone thought the combination (Comb) would provide the best outcome.

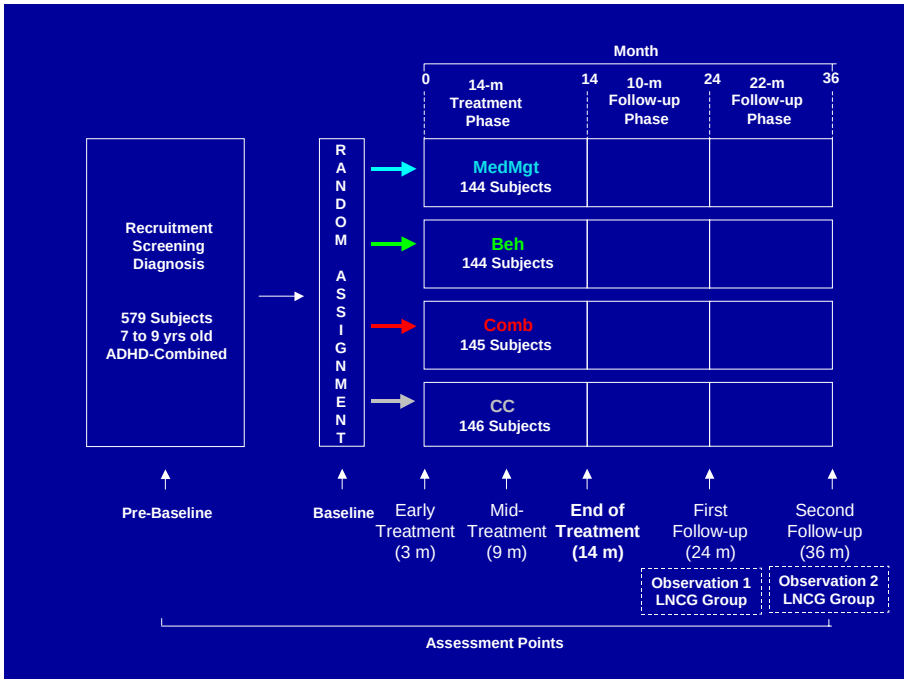
In the MTA planning phase, there was much discussion about a community comparison (CC) group. One of the concerns was generated by the expectation of large variance in this group due to the variety of treatments. If this occurred, then any comparison to this group would suffer. However, the primary hypotheses (i.e., MedMgt vs Beh; MedMgt vs Comb and Beh vs Comb) did not include the CC group. We also debated whether to use some of the limited resources of the study for this fourth group, which would require a decrease the sample size or treatment duration in the three other treatment groups, and we were aware of that this factor would operate and would make it more difficult for any of the planned comparisons of the study to reach statistical significance. But we decided to include the CC group anyway, because we expected other benefits of having this group included.

First, we were concerned about changes in the use of medication in the community, and the CC group would allow us to document this. Since we did not deliver the care in the CC group, we had to document what was provided. In the context of the MTA, we developed the Services for Children and Adolescents-Parent Interview (SCAPI). Information on the development of this instrument was published recently by Jensen et al. (2004) and Hoagwood et al. (2004). The SCAPI was intended to evaluate the treatment that was provided in the usual or natural environment of the home and school settings.

Second, we proposed an evaluation of the relative effects of two intensive, state-of-the-art interventions (MedMgt and Beh), but we did not have an untreated or placebo control group. In fact, some thought it would be unethical to withhold treatments since there were strong empirical bases of efficacy and effectiveness for both stimulant medication and behavior modification. We expected that some variation of these

treatments would be applied in the usual care or CC group, but perhaps with less intensity than in our state-of-the-art treatments.

To summarize the design of the MTA, we used four randomly assigned groups – Medication Management (MedMgt), Behavior Modification (Beh), the Combination (Comb), and Community Comparison (CC). For this randomized clinical trial, we developed a broad assessment battery that was initially applied at baseline and then at an early point (3 months), at a midpoint (9 months), and at the endpoint (14 months) of the treatment phase of the study.



Slide 1

For this basic study design, we performed a statistical power analysis based on our desire to detect an effect size (the difference between means divided by the expected standard deviation for comparison of any of the two conditions) that would be equal to or greater than 0.40 (see MTA Group, 1999). This is considered to be a small to moderate effect size (Cohen, 1988). Based on this power analysis, we determined that we must assign 144 children to each of the four treatment conditions. Thus, each of the six sites was given the charge to recruit 96 children between 7 and 9 years old who met the DSM-IV criteria for ADHD-Combined Type. From 1995 to 1997, we recruited and assigned 579 cases to the four conditions: MedMgt (n = 144), Beh (n = 144), Comb (n = 145), and CC (n = 146).

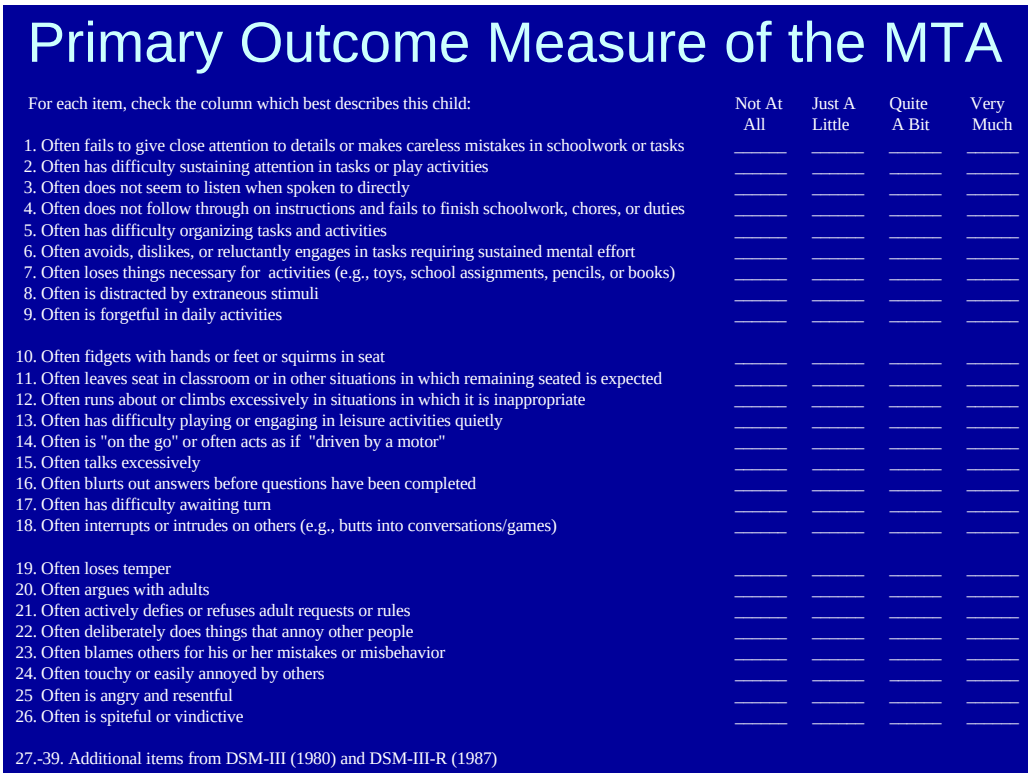
After the initial 14-month treatment phase, the MTA was continued as an observational follow-up study. The first follow-up assessments were at the 24-month and 36-month assessment points, and we are continuing now into year 10. At the 24-month assessment, we also recruited a local normative control group (LNCG) of classmates of the children with ADHD. This is a valuable group that we are now using to evaluate the

degree of normalization of the clinical groups with treatment or with developmental course (over time).

A wonderful statistician, Helena Kraemer from Stanford, introduced us to design and analysis of Randomized Clinical Trials (RCTs). Many of us had never conducted an RCT before, but we followed her good advice to perform a formal power analysis before initiating the recruitment, to use random effects regression to evaluate change over time so we could tolerate missing data, and to decide upon a limited set of simple and specific questions as the primary hypotheses of the study.

We proposed three sets of comparisons. First, we proposed to compare the two modalities (MedMgt versus Beh) to evaluate the relative efficacy of these two state-of-the-art treatments. The advocates of these two treatments agreed to abide by the results. Second, we proposed tests to determine whether the combination (Comb) was actually better than either of the two modalities provided individually, as almost everyone expected (Comb versus MedMgt and Comb versus Beh). The third set involved the controversial group, community comparison. We compared each of the treatment conditions (MedMgt, Beh, and Comb) to a common control (the CC group).

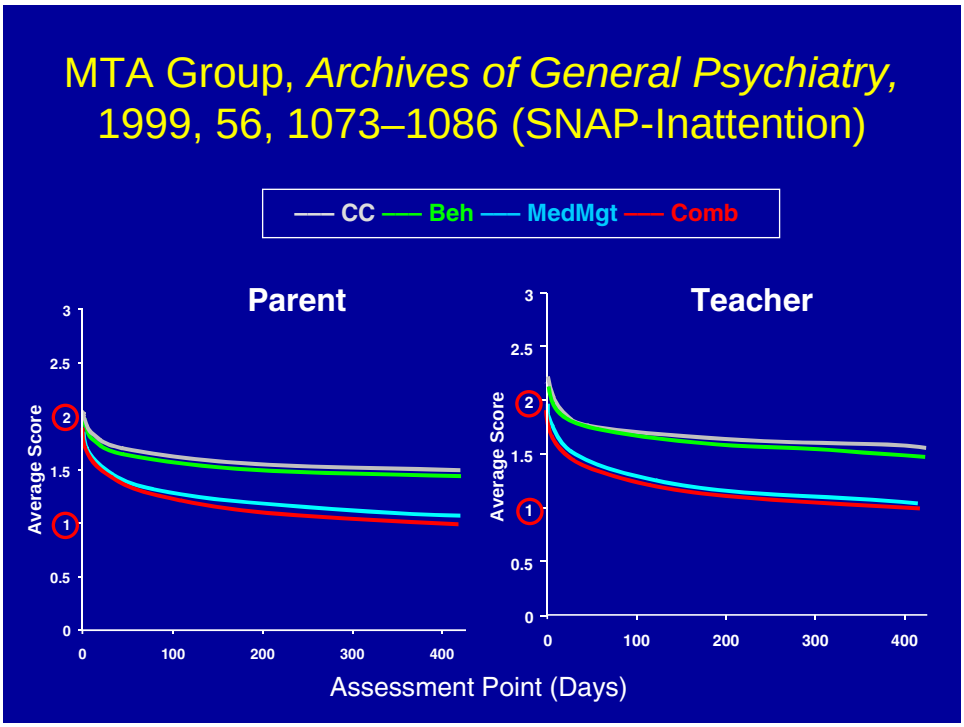
The first comparison was a head-to-head comparison of two unimodal conditions that had strong advocates participating in the design, implementation, and analysis of the RCT. For this comparison, we had debates about many issues – whether to use a single outcome measure, some composite of the large assessment battery, or even a measure of Clinical Global Improvement (CGI), an overall subjective evaluation that was typically recommended at that time by the FDA for evaluating medication trials. In the MTA battery, we had 6 domains of outcome and 19 measures. We decided that a primary outcome measure would be severity of ADHD symptoms that was assessed by rating scales that included the DSM-IV symptoms of ADHD as items. One of these was the SNAP questionnaire (see slide 2), which consists of 18 items grouped by the two domains of symptoms of ADHD (inattention and hyperactivity/impulsivity), plus the eight symptoms of Oppositional Defiant Disorder (ODD).



Slide 2

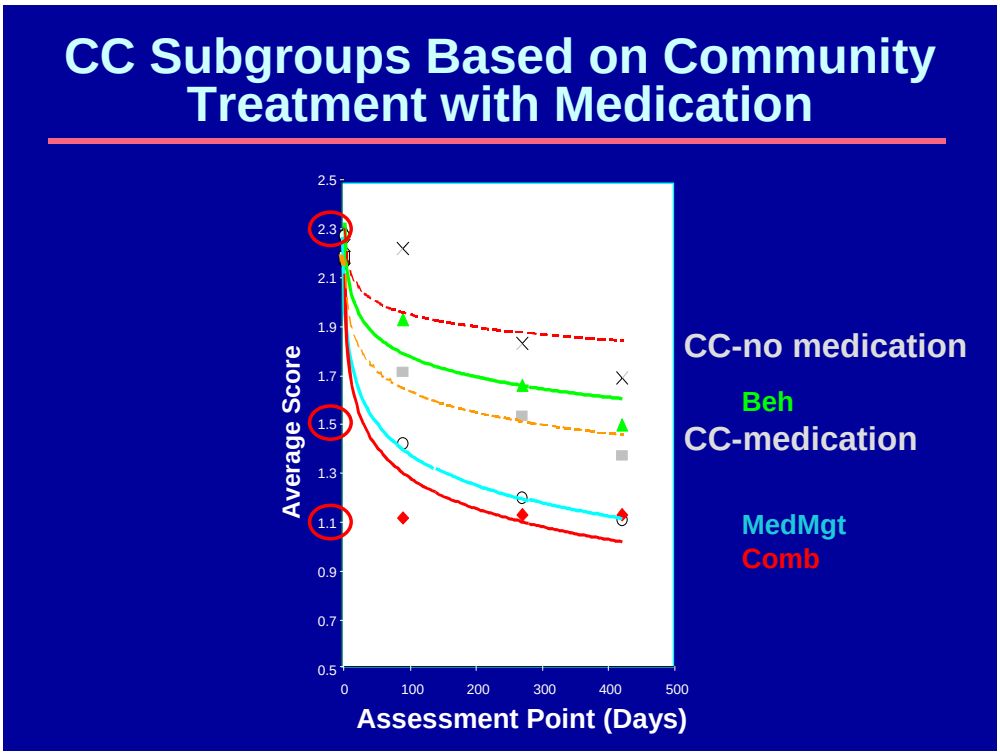
The SNAP was completed by two sources (parents and teachers), who rated each item as being present “not at all,” “just a little,” “quite a bit,” or “very much.” These ratings are scored as 0, 1, 2, or 3, and then averaged for each domain. The average rating of 2.0 is considered high and usually associated with significant impairment in most children so rated. Most of the children entering the MTA had an average symptom rating around 2.0 at baseline. An intervention that reduced the average symptom rating to 1.0 would be considered highly effective, since this level of symptom severity would characterize near-normal behavior.

We published the primary results of the randomized clinical trial phase of the MTA in the *Archives of General Psychiatry* (MTA Group, 1999), which are summarized in slide 3. The first planned comparison – MedMgt versus Beh – produced a clear finding. Even though we used an extraordinarily intensive behavioral treatment, the reduction of severity of ADHD symptoms was less for the Beh group than the MedMgt group. The second planned comparison produced a surprising finding: the medication alone (MedMgt) and the combined (Comb) groups did not differ. This was particularly shocking to the architects of the comprehensive behavior modification intervention used in the MTA. This finding provided the type of rude awakening that one often gets from a randomized clinical trial -- an unexpected result that requires re-evaluation of assumptions about treatment. The third set of planned comparisons involved the community comparison (CC) group. The CC group also improved from baseline to the 14-month assessment, and the magnitude of this improvement was about the same as for the Beh group. However, the improvement was significantly less than for either of the medication groups (MedMgt and Comb).



Slide 3

The SCAPI revealed that most of the children in the CC group received medication. For those that did get medication in the community (about two-thirds of the group), the characteristics of the medication treatment differed from that provided by the MTA algorithm. The percent of time on medication was lower: the total daily dose differed dramatically (19 mg vs. 33 mg per day), and the months on medication differed (about 5.5 months vs. almost 10 months) during the 14-month treatment phase). As shown in slide 4, this subgroup of the CC group had somewhat better outcome that the Beh group, while the subgroup of the CC group (about one-third) that did not get medication in the community was somewhat worse. Apparently, these differences in state-of-the-art medication management (in the MedMgt and Comb groups) and the community standard (in the CC group) resulted in a medium-to-large significant difference in outcome at the 14-month assessment point.



Slide 4

We now can address the questions that concerned us when we debated the use of the treatment as usual or community comparison (CC) group. First, was there increased variance in the CC group? In a secondary analysis (Swanson et al., 2001), we described the means and standard deviations (SDs) of ratings for the four treatment groups, separately for the parent and teacher ratings of Inattention, Hyperactivity/Impulsivity, and ODD (see slide 5). On these outcome measures, our fear was unfounded: the SD of the CC group was not higher than the SD of the other groups (MedMgt, Beh, or Comb).

Increased Variance in the CC Group?

	Parent				Teacher			
Column No.	1	2	3	4	5	6	7	8
Domain	$Inatt_P$	H/Imp_P	O/D_P	$SNAP\text{-}IV_P$	$Inatt_T$	H/Imp_T	O/D_T	$SNAP\text{-}IV_T$
Section A: mean ratings[a]								
Comb	1.08	0.88	0.80	**0.92**	1.16	0.76	0.63	**0.85**
MedMgt	1.18	0.96	0.99	**1.04**	1.28	0.96	0.80	**1.00**
Beh	1.40	1.21	1.05	**1.22**	1.51	1.16	1.00	**1.23**
CC	1.49	1.34	1.11	**1.32**	1.47	1.21	1.01	**1.22**
Section B: SD of ratings[b]								
Comb	0.67	0.64	0.67	**0.57**	0.77	0.67	0.70	**0.58**
MedMgt	0.76	0.69	0.77	**0.65**	0.84	0.80	0.81	**0.72**
Beh	0.67	0.70	0.74	**0.60**	0.80	0.81	0.82	**0.70**
CC	0.70	0.71	0.66	**0.58**	0.83	0.82	0.83	**0.66**
Section C: ES[c]								
Medication algorithm	0.45***	0.52***	0.25*	**0.48*****	0.33**	0.42***	0.36**	**0.45*****
Multimodal superiority	0.14	0.11	0.26*	**0.20**	0.15	0.27*	0.23	**0.23**
Psychosocial substitution	0.13	0.18	0.08	**0.16**	0.05	0.07	0.00	**0.01**

SNAP-ADHD Ratings

2 Sources (Parent and Teacher) for 2 Dimensions (Inatt and H/I)

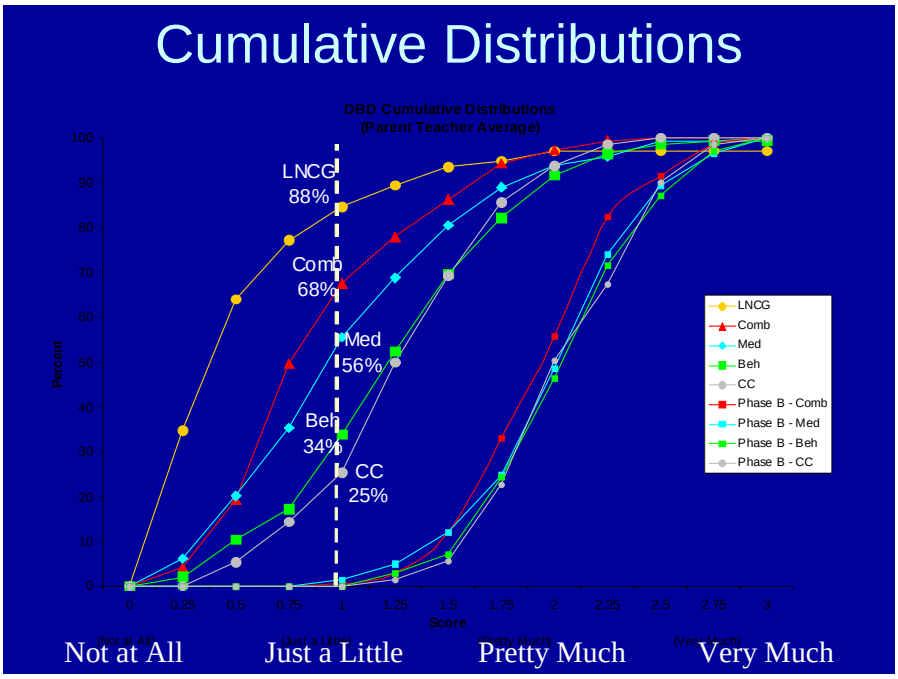
Slide 5

We also addressed the question about sensitivity: what effect size could be detected with the number of subjects assigned to the treatment groups (n = 144 to 146)? In these analyses we used a set of orthogonal comparisons, which address the hypothesis of Multimodality Superiority (Comb vs MedMgt), the MTA Medication Algorithm (Comb+MedMgt vs Beh+CC), and Psychosocial Substitution (Beh vs CC).

As shown in slide 5, for these comparisons, effect sizes as low as 0.26 were statistically significant. This shows that we exceeded our goal of detecting an effect size of at least 0.40. In addition, we increased the sensitivity of our comparisons by developing a composite measure of overall SNAP ratings, based on the average across sources (parent and teacher) and domains (Inattention, Hyperactivity/Impulsivity, and ODD), which increased the precision of measurement, as expected (Spearman, 1919; Brown, 1919), over the individual ratings for each domain and source. In a secondary analysis using this composite to evaluate the hypothesized superiority of multimodal treatment, we were able to detect an effect size of 0.26 for the comparison of the Comb and MedMgt groups, and both were significantly superior to the CC group (reflecting "treatment as usual"). However, the contrast of the intensive behavior modification intervention to "treatment as usual," which we labeled the Psychosocial Substitution effect since such intensive intervention is not typically available in the community, produced smaller effect sizes (< 0.20) that were not statistically significant with our sample size (n = 144 to 146 per group).

We also developed a method for comparisons based on a qualitative measure of success based on a discrete cutoff on a quantitative measure (the average rating across domains and sources), with successful treatment defined by a composite SNAP rating less than or equal to 1.0 (see Swanson et al., 2001). The local normative control

group (LNCG) was useful in determining the adequacy of this cutoff. The cumulative distributions shown in slide 6 show the status of the groups at baseline and at the end of the treatment phase (i.e., at the 14-month assessment). For all groups, the entire distribution shifted to the left, reflecting improvement. The percentage of cases with "successful" treatment or who approached normalization varied across the treatment groups. How many children reach "normalization" from high scores around 2.0 to scores less than 1.0? As shown in slide 6, successful treatment was achieved in only 25 percent in the CC group who received "treatment as usual." The percentages with evidence of successful treatment were higher in the other groups: 34 percent in Beh, 56 percent in MedMgt, and 68 percent in Comb. We used a receiver operating characteristic curve analysis that provided comparisons of each of the three MTA treatment groups to the CC group at each of the six sites of the MTA. This analysis revealed interesting site differences.



Slide 6

In the MTA medication protocol, three-times-a-day dosing with an immediate-release formulation of methylphenidate was used. The rates of successful treatment established this as a "gold standard" for the treatment of ADHD. Another way to deliver the same general pattern of medication is with a once-a-day, controlled-release formulation rather than multiple doses per day of immediate-release medication. Since the MTA, pharmaceutical companies have developed improved controlled release formulations of stimulant medications, for both methylphenidate (e.g., Concerta) and amphetamine (e.g., Adderall XR). These formulations achieve durations of action that approach 12 hours, which match that of the "gold standard" established by the MTA with three-times-a-day dosing with immediate release methylphenidate. As a result, most children with ADHD who once were treated with immediate-release formulations of stimulant medications now take a controlled-release formulation.

We have reported the effects of assigned treatment (using the intent-to-treat analysis framework of a randomized clinical trial) as well as actual treatment (documented by the SCAPI) at the 24-month assessment (MTA Group, 2004). Using naturalistic subgroups based on medication status at each assessment point, we considered adherence to assigned treatment. For example, even in the treatment phase of the study, not all children in the behavioral group were off medication – about 25 percent went on medication in the community anyway – and not all the children in the medication management or the combined groups took medication (about 85 percent did). The analyses at the first followup (at the 24-month assessment) reveal that the outcomes of the assigned treatment groups are converging. This may be a consequence of the rates of treatment with stimulant medication converging over time. In the groups assigned initially to the MegMgt and Comb treatments, about 65 percent are on medication 3 years later, but at this point about 45 percent of the group assigned to the Beh treatment now take medication, also. The community comparison group stayed about the same over time – about 60 percent of those children assigned to the CC condition have been treated with medication every year since the first assessment.

We have completed the analyses of outcome through the 36-month assessment, using new analysis procedures (i.e., propensity score analysis to adjust for self-selection of actual treatment and growth mixture modeling to evaluate heterogeneity of trajectories of outcome over time). The patterns of treatment over time as well as outcome over time were presented and discussed, but the papers describing the results are now under review, and thus the findings cannot be presented here.

Commentary on Case Study #2

Julie Magno Zito, Ph.D., University of Maryland

Dr. Zito is a pharmaco-epidemiologist and associate professor of pharmacy and psychiatry at the University of Maryland. Her most recent research, focusing in children and adolescents, has received extensive media attention. Before she came to the University of Maryland, she was at the New York State Office of Mental Health at the Nathan Kline Institute, where she authored the 1994 textbook, Psychotherapeutic Drug Monitoring. Dr. Zito has been active in the leadership of the Medical Care section of the American Public Health Association.

What do we know from the MTA study community care comparisons? To review the topic briefly, I suggest we look at the goals as Dr. Weinstein has already described them, because the goals of the trial drive the design and the evaluation. Secondly, we should consider some of the scientific issues that we might debate and that some investigators have challenged in regard to the design, the study population, the interventions used, the comparisons made, and the interpretations (which might differ a lot depending on what specialties we come from and our expectations for this treatment area). While other areas like cancer are very clear, psychiatry is a most challenging area in which to bring groups of people together to build consensus around the definition of improvement and choice of long-term goals.

The goals of the MTA investigators were to determine if a 14-month use of medication and behavioral management were comparable in outcomes, whether the combination would be more beneficial than either one alone, and whether the study interventions would be better than routine community care.

This trial is a hybrid of efficacy and effectiveness, because the investigators started from the premise that there is no need to demonstrate whether stimulants work or not because we know they do. Dr. Weinstein alluded to the 300-odd review papers and the 4,000 papers to show that ADHD is the best-studied child question that goes back to 1937 and Benzedrine. This wealth of knowledge sits in stark contrast to many other pediatric drug questions about which we know very little; this is one pharmacologic question in which we happen to know a lot in regard to children because it has been used continuously since the 1950s. The behavioral interventions are well established, so there is an evidence base for the selected behavioral training.

Regarding the scientific design, there were four groups, one of which was medication alone (MED) – largely methylphenidate; in a few instances a child who failed methylphenidate might have been put on a different stimulant or some other medication. All of these were managed by a protocolized approach to medication management, and its results are indicated in terms of the average daily dose at the end of 14 months. It produced a fairly intensive average daily dose of methylphenidate, which might be one way of assuring the desired outcomes. The behavior arm is a "Cadillac approach." The psychologists argue, "It is not that our treatments do not work, it is that providers do not know how to use them properly." This is a case of "we are going to use them properly in order to rule out inadequate exposure to the intervention." Then there is group 3 with combined medication and behavioral treatment and lastly the group that represents usual care.

The MTA used randomized assignment. But I come from pharmacology where we do not just want randomization; we want double blinding. We want placebos and sham procedures and things like that to reduce bias. In the MTA, there is a preliminary double blind phase with a fancy algorithm for dosing methylphenidate to attempt to optimize dose and to rule out placebo response. However, the primary outcome assessments are mostly unblinded, including parent ratings and teacher ratings, with the blinded ratings restricted to a classroom observer. Some methodologists in psychology complained about not having objective assessment because a lot of subjectivity comes with a lack of blinding. If the parents came into the trial with a positive disposition toward medication use and they see their child is receiving medication, then they could be exaggerating the results. The same is true of teachers; we know that there are many teachers who really like these medications. Or the opposite may apply to negatively disposed parents and teachers. Whatever the influence of their viewpoints, we do not know their effect on the study results and the lack of blinding is troubling.

Finally, there is the issue of whether there is a control group to compare with the behavioral intervention. All of these children will be in care for 14 months. If we had a

condition in which merely coming back regularly could be the control group, we might be able to parse out whether minimal continuous care alone delivers a positive response.

Is the study population representative of children in treatment? They screened 4,541 and randomized 579 (12.8 percent); there were some substantial exclusions. Doctors in the community treat children who have multiple physical problems and multiple mental health problems, but those children are not represented among MTA participants. We can expect better outcomes in this trial because the participants are not the "usual" community-based population with its many multiple comorbidities.

The investigators should be applauded for conducting the medication protocol in a systematic way so that children who might be managed on 5 mg/day rather than 10 mg/day had a chance to receive an optimized treatment. Nevertheless, the algorithm for the MED group resulted in substantial dosage – about 38 mg/day exposure. The behavioral treatment was a "Cadillac," an impressive range of parent training and individual therapy for families, group therapy, etc. Russell Barkley (Barkley, 2005) challenged this issue and argued that behavioral training is empirically supported but not theoretically sound. That issue is beyond my expertise. Understanding ADHD etiology and the rationale for various psychosocial therapeutic interventions is a vast domain in itself! On the other hand, the combined therapy participants, who received behavior management and medications, got by with about 20 percent less medication per day (averaging 31.2 mg/day), which, over the course of several years, could add up to a lot of milligrams in the future of each child.

The main comparisons in the study show that symptoms changed in the directions that investigators were expecting – parents and teachers reported that inattention improved by 14 months and parents (alone) reported that hyperactivity and impulsivity also improved by 14 months. Unfortunately, the classroom observers who were blinded to the intervention assignment group did not find these symptom changes to be significantly different in comparisons of the medication arm with community care, the behavioral treatment arm with community care, and combination therapy with community care. This is a big limitation to the conclusions that can be drawn. Other important areas, such as social and academic measures of functioning, showed no changes. The data are weak, at best, to support medication management over the other arms of this study.

What about interpreting the findings? The MTA research team includes many of the most expert people in child psychiatry and behavioral disorders in the United States, so their influence is great and the pediatric and the child psychiatric communities welcomed this study. The impact on clinicians and pediatricians in practice and on future research is likely to be high. The interpretation they are likely to make is that medication is essential to treat children with ADHD. Even though the authors talk about some gains being made for behavioral intervention, this "Cadillac" of treatment has little chance of being practiced in our communities for a number of reasons, not the least of which is that parents of ADHD children are frequently people who are already enormously burdened – in part because of caregiver burden, as a result of socioeconomic status, or because many are one-parent families. The idea that

participants will come back for psychosocial therapy visits on a regular basis is very difficult, not to mention the issue of who will pay for it. Those bar graphs do not compel you to say that behavioral intervention is needed. In addition, a Swedish investigator conducted a traditional randomized double-blind controlled study of dextroamphetamine against placebo (Gillberg et al., 1997). Participants were treated for 15 months, and the authors found the medications to be effective across a 15-month period. That study cost a great deal less money than the MTA.

Here are some ethical issues to think about. Behavioral intervention is highly unlikely to be used in the community because of cost and time. Weisz has already pointed out to us that psychosocial interventions done under research settings are far better able to produce positive outcomes than those done in the community (Weisz, Weiss, & Donenberg, 1992; Weisz & Jensen, 1999). The Fort Bragg study has shown that more is not necessarily better in terms of providing a continuum of care of services for children with mental health problems (Bickman, Summer, & Noser, 1997).

We do have some dilemmas here, and we attack these problems with our research heads but not with our community-minded heads. We have various perspectives that are not always in sync (parent, teacher, clinician researcher, and clinician practitioner), each looking somewhat differently at the risks and benefits of intensive drug therapy in terms of long-term risk to developing hearts, livers, kidneys, and brains, and at effectiveness in terms of tolerability, adherence satisfaction, and improved functioning.

The community care arm was critical in this study. It was not protocolized community care; it simply recorded the experience of children randomized to that arm. Therein lies the rub – these people volunteered to be part of a study by responding to advertisements. How representative are responders to an advertisement who are willing to come back for more than 3 years? It is laudable that they would be willing to be in a community arm and to be monitored for that length of time. Their medication regimen was about 22.6 mg/day, which is 40 percent lower than the medication management arm. Some were only on medication for 5 months a year, so it is difficult to know how much time a child needs to be on medication and whether continuous intensive medication over the long haul will prove to be our best practice in this area. The take-home message is that significantly better medication group response at 14 months changed to a 50 percent drop in effect size at 24 months and was not different across the groups at 36 months.

In conclusion, the MTA is technically impressive – with 579 participants it is the largest study of children with ADHD. Everyone in the field was energized and excited to see that this study was feasible and completed. Its complexity boggles the mind. Completion rates were extraordinary. The researchers are to be congratulated for doing a great deal of work. But the results are unexpectedly modest. I appreciate the propensity score approaches and subgroup analysis, but there is a lot of effort still needed to find out more about why the primary interventions did not work as well as expected.

Besides documenting outcomes of community care under the auspices of a randomized trial, as was done here, we might start looking to observational designs to assess what is going on with ADHD children. We have large cohorts of children in community care around the country. For example, Kaiser Permanente outpatient settings have a computerized information system in which there is the opportunity to evaluate each child; maybe we could add three questions at the end of selected visits to ask how they are doing, if there has been any change, or if side effects led to drug discontinuation or switching. These are simple approaches but it might be all we need to do to establish effectiveness and safety in the community-based population after a drug is marketed. Comparison arms would include placebo and standard therapy. The result is long-term benefit to risk assessment in the children we are likely to encounter in community care.

I like pragmatic trials – large simple trials in community populations – as an approach to evaluate marketed medications. Of course, we have to keep placebo trials because we have the drug development period (pre-marketing) to consider, and every new product has to have that kind of assessment to decide if it is worth marketing. It would help to have standard care in the community arm so that the relative worth of the newer product can be assessed. Augmentation trials may be useful for combinations, but a lot depends on the particular drugs and/or medications, the particular behaviors, and how easy it is to get individuals on and off medications. Lastly, we need to examine specific subgroups of stimulant users, particularly those receiving complex combinations of psychotropic medications (polypharmacy).

References

Barkley, RA (2005). Commentary on the multimodal treatment study of children with ADHD. *Journal of Abnormal Child Psychology, 28,* 595-599.

Bickman, L, Summer, WT, & Noser, K (1997). Comparative outcomes of emotionally disturbed children and adolescents in a system of services and usual care. *Psychiatric Services, 48,* 1543-1548.

Gillberg, C, Melander, H, von Knorring, A, Janols, L, Thernlund, G, Hagglof, B, et al. (1997). Long-term stimulant treatment of children with attention-deficit hyperactivity disorder symptoms: a randomized, double-blind, placebo-controlled trial. *Archives of General Psychiatry, 54,* 857-864.

Weisz, JR & Jensen, PS (1999). Efficacy and effectiveness of child and adolescent psychotherapy and pharmacotherapy. *Mental Health Services Research, 1,* 125-157.

Weisz, JR, Weiss, B, & Donenberg, GR (1992). The lab versus the clinic: effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578-1585.

Panel Discussion of Case Study #2

Constantine Frangakis, Ph.D., Johns Hopkins University

Dr. Frangakis is an associate professor in the department of biostatistics at the Johns Hopkins Bloomberg School of Public Health. His research covers a remarkable diversity of topics, ranging from HIV to Alzheimer's to transportation safety data.

This study was a well designed and executed trial. It is important to study more intensively what is going on in the usual care group. We saw from this trial, for example, that the usual care group had a number of children who went on to use medication. That was possibly one reason why the medication-management-alone group was not so well distinguishable from the community care group.

Why is this important to study before we start the trial? For example, suppose we can identify two subpopulations in the usual care group – one group of children that uses no or very little drugs ("non-users") but among those who use it there is a great effect, and one other group who uses the drugs but there is no effect. If we could know which children are in these two groups or if we could have useful predictors of these groups, what would we do from the design perspective? If we know who is in the group that uses the drug but experiences no effect, then we would not give the drug to these people because we would know from the beginning that it would not be effective. In those people we would try something else. On the other hand, in the other children who are unlikely to use the drug but who experience a big effect when they do use it, we would target our intervention to these children to increase the utilization of the treatment management.

By studying the usual care group before the trial begins, we can have some customization of where to deliver what treatment in order to have a more efficacious result. We do need methods to identify these groups to tease out the effect of drug utilization.

For followup of these children, we now have four groups, all of which are usual care groups but differ originally in being assigned to different treatment arms in the trial. In these cases, in order to do an intention-to-treat comparison, we are fine. But as Dr. Swanson said, after you stop delivering the treatments and you follow up, you need methods to tease out the facts of what people actually do. In both of these issues, namely, the subgroups within a usual care arm, and the follow-up of all the groups after the trial, I see the need to develop additional methods of analysis. We need to identify instrumental variables that are present in these subgroups in order to find more specific causes and effects of treatment.

Paula D. Riggs, M.D., University of Colorado Health Sciences Center

Dr. Riggs is an Associate Professor in the Department of Psychiatry at the University of Colorado at Denver and the Health Sciences Center (UCDHSC). She is the Director of Psychiatric Services for Adolescents at the UCDHSC-affiliated Addiction Research and Treatment Services (ARTS) Adolescent Treatment Programs. Dr. Riggs' research has focused on the development and testing of effective pharmaco-therapy and behavioral treatment interventions in adolescents with substance use disorders and co-occurring psychiatric disorders.

I would like to confess my biases before commenting on the design and methodologic questions raised by the MTA study, so well formulated by Dr. Swanson.

When I think about design and methodological approach of a clinical trial (including the appropriateness of a treatment as usual [TAU] arm), I first think about where the state of the science is with regard to the question(s) that are being addressed. A thorough review of previous research is needed to understanding the leading edge of scientific knowledge in a particular area of inquiry, which in turn informs the next questions that need to be addressed to further extend research in this area. Formulating the next questions or hypotheses informs/guides the appropriate design, sample characteristics, methodological approach, setting, type of control/comparison group, and outcome measures and statistical power necessary to address these questions, scientifically.

The decisional balance with regard to study design must weigh scientific rigor against practical implementation, subject recruitment/feasible sample size, and cost issues. The necessary sample size for adequate statistical power to address study questions is influenced by the degree of control that is necessary over sources of unmeasured variability, impacting key outcome measures that may confound interpretation of results. Consideration of this last point is critical in determining whether TAU may be an appropriate setting or comparison arm in the study design.

TAU may introduce a source of (unmeasured or unmeasurable) variability that impacts the study's primary outcome measure(s), which would require a much larger sample size to ensure sufficient statistical power to address the primary study questions. If the sample size necessary to overcome TAU variability is not feasible in terms of subject ascertainment, cost, and timeline considerations, then TAU may not be an appropriate treatment arm or setting for this particular study. In such cases, a comparison group or treatment arm with less "background noise" and tighter control over sources of variability than TAU may need to be considered.

This brings us back to stage of science. For example, if there have been no previous randomized controlled trials supporting the efficacy of a particular intervention, this tells us that to propose a multisite effectiveness trial in real world (TAU) settings is premature in terms of the stage of science. Randomized controlled trials evaluating the initial efficacy of a treatment intervention are most often conducted in a single site and require fairly tight control over sources of background variability in order to feasibly recruit an adequate sample size to achieve the statistical power necessary to address efficacy and estimate effect size. In such studies, TAU is often not a feasible setting. TAU may be appropriate as a comparison group if the study question is whether Treatment A is superior to TAU. The difficulty here is "what is TAU" and/or how TAU can be standardized to more broadly represent TAU. Researchers have proposed that manualized TAU is a virtual requirement in such cases (Carroll et al. 2000). But if manualized and standardized, is it still TAU?

Choosing TAU as a research setting or platform is most often appropriate once the question of the intervention's efficacy has been determined. For example if three randomized controlled trials independently conducted at three different sites by different investigators with similar, but regionally different, samples have demonstrated the efficacy of treatment intervention "A" compared to placebo or TAU or other comparison group, then the question of the intervention's efficacy under fairly tightly controlled

conditions has been sufficiently addressed. The next question in this case is whether treatment "A" can be feasibly implemented and shown to be effective in multiple real-world settings with a broad range of real-world patients. In effectiveness trials, TAU is perhaps the most appropriate comparison group or treatment setting. Exceptions to this sequencing are in cases of hybrid efficacy/effectiveness trials (beyond the scope of discussion here, but sometimes appropriate when adequate sample size cannot be feasibly recruited by a single site for initial efficacy testing).

When applied to the MTA trial, these principles would affirm the appropriateness of TAU as a comparison arm. The efficacy of methylphenidate had been demonstrated in a number of well-designed controlled trials prior to the MTA study. A multisite effectiveness trial was the appropriate next step to extend the state of the science in this area. Another strength of the study design is the excellent biostatistical expertise and consideration given to adequate statistical power from the outset – during the initial design of the study. It was very impressive that *a priori* consideration was given to powering the study to detect the lower limit of a clinically significant effect size (0.4 being quite reasonable) in order to guard against a potentially underpowered negative study with uninterpretable results.

However, I did wonder whether the principal biostatistician understood that 19 "primary" outcome measures were going to be selected and how that may have impacted the study's power analysis! In other words, did you really have adequate power to detect a lower limit of 0.4 effect size for all 19 variables? It's also not clear from the presentation or the published manuscripts of main study findings to what extent fixed versus random effects across participating sites were considered in the initial power estimates for the study.

Another strength of the study was the restriction of age range. This is important as a way of reducing variability in the sample – electing not to mix children with later adolescents in terms of developmental variability that may impact ADHD outcomes. Moreover, generalizability of results is enhanced by the decision not to have overly restrictive inclusion criteria – for example, including children with other co-occurring disorders. However, it was not clear to me how many of those with comorbidity might have received treatment (or not) for these disorders and whether that varied across sites, and how this might have impacted interpretation of results.

The fact that you had no baseline differences on important outcome variables is an important (although somewhat serendipitous) strength of the study. I also thought that the design choices related to the single blind and no placebo treatment arm were justified in an effectiveness trial of a medication, for which the efficacy has been clearly established scientifically. Moreover, the placebo effect size had also been well established prior to the MTA trial.

The main criticism of the study is related to an aspect that could probably not have been anticipated at the time the study was designed – the extent to which the community-based TAU changed during the study timeframe. It appeared that the MTA interventions also "morphed" somewhat over the time-course of the study, thereby

introducing variability that most likely cannot be fully assessed or accounted for in analyses. For instance, it seemed that the taper sequence for the intensive behavioral intervention was not carried out entirely according to *a priori* determined methods.

My most important question related to study design is whether the MTA interventions were offered at no cost to study participants/families and whether those receiving TAU in community treatment settings were charged for treatment services. If so, this would be my most substantive criticism of the study in terms of threats to interpretation of results. If people in the community arm were paying for treatment services and people in the MTA arms were not, this would be an important factor influencing treatment retention and compliance, both of which were important study outcomes for which differences between treatment arms were found and reported.

Lastly, the study result I find most intriguing is related to the neurobiological aspects of ADHD that has implications for future research. You noted that it was puzzling why it appeared that so many children who took but later stopped medications appeared to have persistent treatment effects, given the chronic nature of the illness. You wouldn't have expected the illness to remit during this time period. Why did so many subjects appear to have persistent treatment effects? We have learned from neurodevelopmental research that the children in the MTA sample were being treated during a time of maximal gray-matter pruning but increased myelinization. Could it be that optimal pharmacotherapy for ADHD during this time of development might produce lasting neuroadaptive changes even if medication for ADHD is discontinued – because it impacted grey matter pruning? Or because treatment enabled more optimal learning that then mediated more normalized neurodevelopment during this critical period? To my knowledge this question has not been addressed and supports inclusion of more basic science approaches –including incorporation of assessment of relevant biomarkers and/or conducting neuroimaging studies in a subsample of participants as an aspect of the design of future clinical trials.

In conclusion, MTA was an important study that was designed to address the most appropriate next research questions in the context of the state of the science at the time it was designed and implemented. And, as with all good research studies, results raised important questions that help clarify important methodological issues and directions for future research.

Betty Tai, Ph.D., National Institute on Drug Abuse, NIH

> *Dr. Tai is director of the Center for the Clinical Trials Network at the National Institute on Drug Abuse. Dr. Tai received her Masters degree from the University of Massachusetts and her Ph.D. degree from George Washington University. Her current endeavors focus on translating drug abuse research into drug abuse treatment.*

The MTA study addressed at least two issues involved in treating ADHD children: 1) determining which of the following treatments is superior – medication treatment alone, behavioral treatment alone, or combined medication/behavior treatment, and 2)

determining how the treatment was delivered. A routine community care comparison group was chosen for both purposes.

Based on our discussions here today, we have observed that in conducting a pragmatic trial, the tendency is to use usual or routine care as the comparison group to increase the relevancy of the tested therapies for practice. For this multisite trial, three sites on the East Coast and three sites on the West Coast were chosen. These sites were all located in areas with highly sophisticated pediatric psychiatric practices. One question arising from the site selection is whether the routine community care in these areas is representative of a national norm. In answering this question, we should ask how "routine" is this routine community care? Does the routine community care in a small town in rural America differ greatly from care that is available in one of the tested regions due to constraints in resources?

Additionally, in contemplating the use of the study results to inform practitioners and/or to impact practice, the transferability and sustainability of the treatment becomes important. The MTA study demonstrated that routine community care was inferior to the medication and combined medication/behavior treatment approaches. In examining the "Cadillac" treatments studied in this trial, one wonders how many clinicians nationwide can be trained to deliver the intervention package with the requisite intensity involved in these treatments. Are these treatments affordable? The cost of training the therapists and delivering the therapy are important practical considerations for dissemination. The "Cadillac" approach may well be too much. We have evidence suggesting that there is a bell-shaped curve for behavioral intervention intensity and its effectiveness; there appears to be a threshold where more is not always more effective.

ADHD can be described as a brain behavior disease, where medication controls the neurobiology deficit of the brain acutely and behavior treatment teaches coping skills that then yield longer-term benefit. If this is the case, then it is difficult to explain why a combined behavior/medication intervention did not yield better treatment outcomes than each of the individual treatments alone. This would be an excellent topic for further research.

Charles Weijer, M.D., Ph.D., FRCPC, University of Western Ontario

> *Dr. Weijer is Canada Research Chair and associate professor of philosophy, medicine, and epidemiology in biostatistics at the University of Western Ontario in London, Ontario. His interests include research ethics and philosophy of science.*

During my talk, I expressed some reservations about hybrid designs like this that have multiple arms and that add a usual care arm onto those multiple arms. I saw some costs associated with it and a couple of benefits, but the costs may outweigh the benefits. What are the benefits of including this usual care arm? I have a couple of alternative scenarios in mind that may help focus that question. One is, what did you learn by having the usual care arm compared to a three-arm trial without a usual care arm? Secondly, what did you learn from a usual care arm that you would not have

learned from a fourth arm that was protocolized but less intensive treatment with methylphenidate? Finally, what did you learn from having a usual care arm that you would not have learned from having a three-arm randomization and then in parallel doing an observation study? Reflection on these questions might help us tease out what we are getting from a usual care arm, given the costs associated with it.

Dr. Swanson's Response to Commentary

If we had not used a community comparison group, how many participants could we have added to the other groups to increase power and what would that increased group size do? The biggest limitation on sample size was the intensive, expensive interventions for the three MTA-treated groups. But if we had used a less expensive, less intensive behavioral treatment (behavioral treatment being the most expensive modality) we would have been criticized for not having been intensive enough in the behavioral group. Since the community care group was not as costly as our intervention groups, we could not have added many participants to the Comb, MedMgt, and Beh groups by eliminating the CC group. Perhaps we could have increased the group size 10-20 percent at most, which would not have increased our power much. After the fact, we can justify our choice in retrospect by noting that we detected a really small difference – effect size of 0.26 – so we did not need any more statistical power.

I like this idea and probably now will think through some of the other protocol possibilities: a protocol of community care that we could standardize so we could learn something from beyond just characterizing it as having a mixture of two different things, either no medication or medication that was accepted. That goes well with Dr. Frangakis' sophisticated conceptualization of these potential people in community care, which we can start looking at with sophisticated techniques – non-compliers who would never take medication even if it was going to work and the non-responders who would try it but for whom medication would not work. This is very interesting to try to understand something about those groups, and we would have to do something else to characterize these theoretical groups of non-compliers and non-responders. The conceptual groups exist but many times you do not have very many subjects (or anyone) in these theoretical groups. In the evaluation of the 36-month assessments and beyond, another statistical expert (Sue Marcus) is advising us on how to use propensity score analysis to investigate these self-selected groups. Also, an expert on growth mixture modeling and latent class analysis (Robert Gibbons) is advising us on how to evaluate and understand natural trajectories over time, which may be different across subgroups, to tease out some of these specific issues.

Dr. Riggs has conducted a clinical trial of adolescents; that is too difficult for me to consider because working with children is difficult enough! This whole idea of looking at the possible impact for early treatment on brain development is truly intriguing and very controversial. An add-on study of the MTA by Jeff Epstein and several colleagues is using fMRI brain imaging of children with parents who have ADHD to evaluate a subset of familial ADHD, which might start to address the questions asked by Dr. Riggs.

My former collaborator and good friend, Xavier Castellanos, is evaluating whether treatment with stimulant medication may accelerate development of brain white matter. This unique hypothesis was generated from a longitudinal study of brain development conducted in the NIMH intramural program, where he worked for many years. He is now continuing this approach in study at New York University, where he is now located. A greater rate of development of white matter with treatment than without treatment with medication might have some impact on resolving a biological factor responsible for some aspects of ADHD, which then may produce a permanent effect. If that is true, it will be a very important finding.

Dr. Tai knows more than I about clinical trials, and this "Cadillac" treatment in the MTA is certainly a weakness of our study. I have no clear idea about what community standards would be outside of the six sites in the MTA. These sites were chosen as centers and they were all in areas where sophisticated treatment of ADHD was available. I referred to one paper on successful treatment defined by a categorical outcome measure of success (Swanson et al., 2001). We performed a Receiver Operating Characteristic (ROC) curve analysis comparing within any site the treatment group differences on this categorical measure of successful treatment, which produced some surprises. The sites that were most involved in developing the behavioral treatments (and probably delivered the most intensive behavioral treatment) did not provide the strongest evidence of positive effects of behavioral treatment. In fact, they showed the opposite effect – combined treatment was worse than medication alone. This was completely counter to our expectations. One of the problems is that the community comparison (CC) groups within the sites may have been tainted by the sites that provide and affect available intensive intervention in the community (i.e., the community standard for behavioral treatment may be much greater in some sites than others). That may have washed out the effect we expected.

One thing that has changed dramatically in our followup is that the MTA results have affected the community treatment of children with medication. This complicates things for scientific evaluation of long-term outcome of children with ADHD now in an observational follow-up study. Almost everyone now treated with stimulant mediations now have coverage for 12 hours a day, because that is the defining characteristic of the new controlled-release formulations of stimulant medications (due to the drug delivery systems in use, a morning administration produces sustained effects that lasts 12 hours). Those products are designed with the MTA medication algorithm as a target. Thus, the community standard is shifting because the MTA affected the pharmaceutical companies to mimic the "gold standard" provided a decade or more ago by the MTA medication algorithm. The widespread use of these new controlled-release formulations of the stimulant medications reduces the variation of treatment regimes in the community.

End of Day One

Case Presentation:
Case Study #3:
Spine Patient Outcomes Research Trial (SPORT)

James N. Weinstein, D.O., M.S.
Dartmouth Medical School

Commentary:

Steven N. Goodman, M.D., Ph.D.
Johns Hopkins University
Dennis O. Dixon, Ph.D.
National Institute of Allergy and Infectious Diseases, NIH
Alex John London, Ph.D.
Carnegie Mellon University
Jon D. Lurie, M.D., M.S.
Dartmouth Medical School

Case Presentation: Case Study #3:
Spine Patient Outcomes Research Trial (SPORT)

James N. Weinstein, D.O., M.S., Dartmouth Medical School

Dr. Weinstein is professor and chair of the Department of Orthopedic Surgery and professor of community and family medicine at Dartmouth Medical School and Senior Member, Center for the evaluative clinical sciences. He is also the principal investigator of the SPORT trial, funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS). Dr. Weinstein founded both the SPINE Center and the first-in-the-nation Center for Shared Decisionmaking at the Dartmouth-Hitchcock Medical Center/Dartmouth Medial School.

SPORT has been an interesting venture, and we are still working on it so I will not be presenting any data. I will provide some background information and how we got to SPORT, and I will talk about the trial design and some of the limitations.

When clinical situations allow for several treatment alternatives, patients need to be empowered with evidence-based information to achieve the right rate for that individual. Clinical trials are indispensable. They continue to be an ordeal – they strain our resources and they protract the moment of truth to excruciating limits. In major medical dilemmas, of which some spine surgery is one, if the alternative is to pay the cost of perpetual uncertainty, have we really any choice?

Spine surgery is interesting, both nationally and internationally. The United States is number 1 in the world in spine surgery. We do not have quite as many orthopedic and neurosurgeons as does Sweden, but on a per capita basis we have many and we do more spine surgery. I came from the University of Iowa to join Dr. Wennberg, who helped me understand the issue of variation and the importance of trying to find the appropriate answer for my patients to questions about the best evidence. As a surgeon, I do not want to perform an operation just because it is available; I want to operate because I offer my patient the best alternative and, given their preferences, the appropriate treatment. Dr. Wennberg and I have written about these issues. In Dr. Zerhouni's NIH Roadmap Initiative, an initiative like SPORT, in which we use patient preference as part of our trial, is a road forward about quality of healthcare and getting patients involved in decisionmaking, given good information – whether that is cancer treatment, tidal volumes on a respirator, or elective procedures like spine surgery where a lot of choice exists.

SPORT is a large "practical clinical trial," really six trials in one. The rationale is that spine surgeons see all kinds of patients. If we are going to set up an expensive infrastructure to study this variation, we wanted to study the major alternatives simultaneously. We are looking at herniated disks, spinal stenosis, and degenerative spondylolisthesis, which are the most common reasons for which spinal surgery occurs.

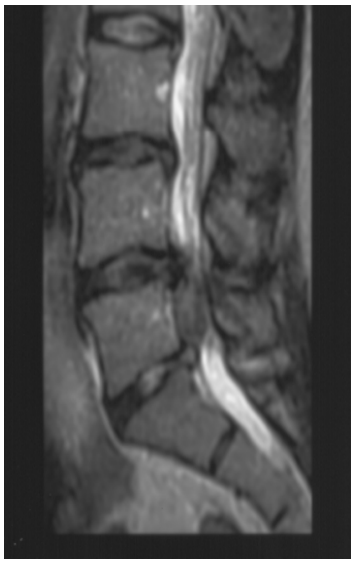
Steve Cummings, who has done a lot of work on musculoskeletal disease at the University of California at San Francisco (UCSF), talks about "FINER" projects – feasible, interesting to the investigator, novel, ethical, and relevant (Cummings:

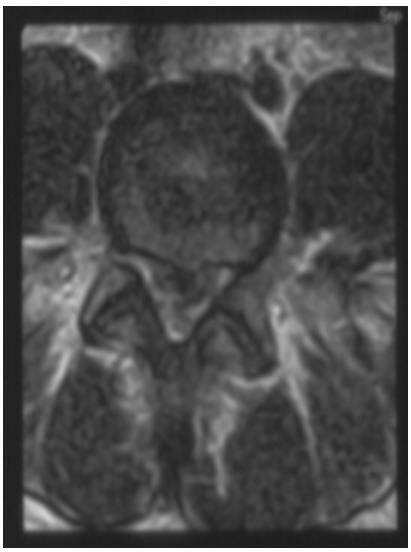
Designing Clinical Research, 1988); all of these elements are true here. SPORT became feasible through 25 years of collaboration with colleagues around the country. It is certainly interesting to me and to many of my colleagues, and the approach to the study design is novel, ethical, and relevant to patients facing these decisions.

Research questions must be well constructed. Our questions compare operative to non-operative patients and look at specific self-reported outcomes, e.g., pain relief and function. Specifically, we are using some of the subscales on the Medical Outcome Short Form 36 Health Survey (SF36) that ask about bodily pain and physical function consistent with studies conducted by Steve Atlas and the group in Maine. We are also using the Oswestry Disability Index (ODI) as a primary outcome measure. Sciatica bothersomeness (questions about whether the individual's leg pain feels better) and satisfaction against expectations are some of our secondary outcomes. One of the interesting parameters in trials is patient expectation prior to treatment and whether those expectations were met and satisfied. Cost effectiveness, which has never been studied prospectively in this field, becomes extremely important from the standpoint of outcome measures.

Is it ethical? In SPORT, investigators have equipoise and patients are fully informed and free to decline. The way in which we informed patients is a model that could be incorporated in other studies. Regarding equipoise, I found an article in the *Bioethics Journal*: "A clinical trial is ethical only if some form of equipoise between treatments being compared obtains…I suggest that while this work is promising, it still has far to go. While equipoise remains the best theory we have of the cognitive justification for clinical trials, it is nonetheless incoherent" (Ashcroft, 1999).

Will the study give the best possible answer? We hope so; if we did not believe so, we would not be doing this study. Some could consider having no treatment for these patients unethical. We are prescriptive in the protocol about non-operative care being physical therapy and medication, etc., but the reality is that people with back/leg pain generally seek relief from anywhere they can. We wanted to make sure we understood what the patients were getting. We have independent monitoring from the DSMB and IRB approval in 11 States.





Here's the spine. Sometimes the spine hurts. This picture is one of the reasons why this trial is so important. The pictures shown above are of a herniated disk. About 20 percent to 30 percent of you (people 50 plus) might have herniated disks on MRI and wouldn't know it. The first question is: why do some people have a herniated disk pressing on the nerves present with symptoms, and other people do not? When we followed up with this patient with a herniated disk, the herniation went away on its own; this scenario occurred two more times in the same patient. But if some herniations resolve on their own with no treatment, why do we perform surgery? Why is one patient experiencing more pain than another?

From a basic science standpoint, we know that form does not follow function; we need both the basic and clinical sciences. The animal models suggest that the animals get better when we simulate herniation and their function does not necessarily correlate with the anatomic findings at all. In the spinal stenosis model, in which nerves are compressed, we also see that the intensity and length of compression will affect the outcome of the nerves. These are important biological phenomena for surgeons to think about when treating their patients.

With MRI and CT-scan, we can now see all kinds of things on imaging studies, but what do they mean? Many people can have imaging findings that do not correlate with their symptoms. This is a patient with spinal stenosis who has a narrow spinal canal. The nerves have nowhere to "live and work" and they are totally compressed compared to this normal area; the nerves are being squeezed. Physical therapy or a drug will not change that squeezing, but what is the natural history and what is the result of surgery? Can nerves recover if they have been squeezed too long?

Scientific equipoise for doing this trial is present, and raises questions at the basic and clinical levels. Low back pain is one of the most widely experienced health problems in the world; about 80 percent of people have it at some time in their life and, on any given

day, about 30 percent of the population experiences back pain. It is second only to the common cold as reasons cited for primary care physician visits or days lost from work. Estimated costs are in excess of $70 billion annually. We know from Dr. Wennberg's work that there is tremendous variation in how low back pain/leg symptoms are treated. However, unlike the hip fracture example in which we all agree on the evidence to treat, here we do not agree and the variations are explained in this case in part by the lack of information. Limited evidence has demonstrated the effectiveness of surgery versus nonsurgical treatment.

Though surgical criteria have theoretically become more stringent, rates of lumbar surgery continue to increase with wide geographic variation, some poor outcomes, and reports of high re-operation rates. Surgical failures are sometimes attributed to poor technique but more often attributed to poor patient selection. These results suggest lack of consensus and lack of clear indications for diagnostic testing for surgery – when the test should be ordered – and the type of surgery being offered – open, closed, fusion, or no fusion. More importantly, patients lack information on the expected benefits and risks, shared decisionmaking, and "informed choice." It is a dilemma for surgeons to avoid inappropriate operations and for nonsurgeons the inappropriate referrals to surgeons.

Here is a map of work that Dr. Wennberg and I have done looking at the change in rates of surgery. In 1996 there is a tremendous increase in spine surgery rates in the United States that we believe, in part, is due to the new technologies available. We are able to do all kinds of things around the spine today, even though it is a complicated structure.

The literature that supports this says that there is only one randomized trial for disk herniations that is quoted often by Weber, and it has poorly defined inclusion/exclusion criteria, lack of conformity of diagnostic images, and lack of physical findings. Spinal stenosis has essentially no RCTs, and degenerative spondylolisthesis has only a few RCTs.

Historically, the disk was first operated on 1934 at Massachusetts General Hospital in Boston by Mixter and Barr, an orthopedic surgeon and a neurosurgeon. They thought they were taking out a tumor but what they took out was part of the disk. It was not until 1978 when a first study was done by Heinrich Weber in Norway; this study is the most quoted study about this surgery in the spine literature today. His outcomes were good; there were no hard endpoints, patients were only randomized after differential decision, and there were no real baseline assessments. The literature criticizes this study, as I am sure they will criticize ours in the future.

Sarpenyer first described spinal stenosis in 1881 from pathologic specimens. Verbiest, a neurosurgeon in Norway, was the first to look at it with myelography in the 1960s. The basic studies suggest that these patients do more poorly being followed for longer periods of time after surgery, but there is really no good information.

SPORT Study Design

SPORT is a multicenter, clinical, randomized, controlled study; one RCT and the other an observational cohort. To my knowledge, this kind of trial has never been conducted in surgical trials. We wanted to answer the question of generalizability, because many surgeons believe that most clinical trials are not generalizable to their practice – some might say, "You can do this at an academic center but it does not apply to me and my patients." We wanted to know what happened to the patients who did not randomize so we could talk about generalizability of the randomized trial arm versus the observational arm. We also wanted to work with patients' preferences by offering informed decisionmaking and informed choice.

Some 86 percent (approximately) of approached patients decided to participate in this trial, about 33 percent randomized, which is very high for a surgical trial. In studies in the late 1980s when I was at Iowa, we wrote out a scenario and hypothetically asked patients, offering them surgery and other options. About 60 percent of those who had a preference wanted surgery and about 40 percent wanted not to have surgery. All patients have the same symptoms – pain, leg symptoms, and MRI images compatible with their diagnosis. These are all people presenting with the same symptoms making their own "informed choice" independent of their symptoms. The effects of preference will be interesting to study, whether they were correct or incorrect in their choice. The randomized patients were the people who just could not decide, and they are the appropriate people to randomize.

This is a pragmatic study of office practice and healthcare effectiveness. We have private practice centers, academic centers, and all kinds of physicians, trying to be as generalizable as possible to represent what is happening in the real world. Strict inclusion criteria for the disk herniation group included physical examination findings, MRI findings, plain-film findings, and symptoms for at least 6 weeks before they could enter the trial; the timeframe was 12 weeks in the stenosis groups. These timeframes were required because of the natural history of some patients getting better without any treatment. We wanted to make sure these patients would remain symptomatic before entering them into the trial; everyone agreed to this approach, given the basic science information and the clinical information that was known.

What is the rationale for the "usual care" control group? From reviews of the *British Medical Journal* on evidence-based medicine for the last several years, I have discovered that there is little evidence to support things we do. Therefore, we had the reality of a minimum of nonoperative care (physical therapy and nonsteroidals) combined with patients going to anyone to get relief of their pain and symptoms. We wanted to be realistic and to capture those things, and designing a protocol in which investigators were blind to everything else would not reflect reality.

When we look at the first data at baseline from our observational cohort, participants are doing what we thought – most are getting anti-inflammatories and physical therapy and a few are using narcotics, especially the disk patients because of their severe pain.

None of us would want to be forced not to take narcotics if we can barely walk, so it is a pragmatic approach to the problem.

For informed consent we focused on the model of "informed choice." Surgeons are biased because we each believe what we do works; therefore, we need to give patients independent information. In SPORT that means giving information across 11 States and approximately 142 physicians, so that the patient is getting at least some core information that is evidence based and unbiased. We accomplished this by using shared decisionmaking videos – hearing from other patients who have or have not had surgery and who have good or bad results, telling the participant who is enrolling in the trial what has happened to them. In the background is the data from the literature showing the expected outcome of surgical versus nonsurgical treatment. We felt this was truly informed consent and truly informed choice, and the doctors also talked to the participants. A total of 98 percent of the participants saw the whole video, which is 53 minutes long. If your back is hurting and you are thinking about surgery, you will watch a long video. The doctors may not like it because it takes away some of their ability to interact with their patients, but we believe we balanced that. In many trials, a video like that would have been a much better way to give information to potential trial participants. We also have long forms that everyone can read, generally at an eighth grade level.

What is the impact of participant and physician intervention preferences on randomized trials? This has recently been written about in *JAMA* in several articles. The conclusion issue is that preferences influence whether patients participate in RCTs, and it was posited that preferences might have decreased the randomization. We had conducted a survey with a set text and had gotten 33 percent; when we actually conducted the trial we also had 33 percent, so it did not seem to effect our randomization. There is little evidence that preferences affect validity.

In monitoring our trial, we talked to the site principal investigators every month and the nurses received calls every month as well. These are the sites, with approximately 2,500 patients total: 33 percent were in the randomized trial and the observational study had 67 percent; 60 percent of participants chose surgery and 40 percent chose nonoperative treatment. When we closed the observational arm, which enrolled a little faster, randomization was not affected and continued at about the same rate. Patients continuing to get their preference and the practice continued, but we did not continue to enroll participants because we had reached our *a priori* sample numbers. Our followups over time were very smooth by groups; so far we have lost very few participants to followup. Disk herniation patients are generally 33 to 55 years old; people with stenosis are generally in their 50s to 90s.

There are several problems with surgical RCTs, including a learning curve. Surgical proficiency is an issue; surgeons argue that they are better surgeons than others and therefore their results are better. The inability to blind participants is also an issue because it is difficult to put a scar on someone and not tell them you did something to them! Patient accrual can be difficult; we were very fortunate to fully enroll this study

although many people said it would never happen. It is difficult to do placebos for surgery.

What we hope to say to our patients is that each individual patient, given useful, evidence-based information, should make the appropriate treatment decision. We will be able to give them better probabilities of outcome based on the best data we have available to help them make that decision; for surgeons that is a great opportunity.

“History is a pack of lies about events that never happened told by people who weren’t there.” “Skepticism, like chastity, should not be relinquished too readily.” I am a bit of a skeptic but I am very proud of this study and hope we will have some information to help patients in difficult situations trying to make the appropriate decisions for themselves.

“Knowledge is power.” We are always learning and will continue to learn and use that information to help others.

Reference

Ashcroft, Richard (1999). Equipoise, Knowledge and Ethics in Clinical Research and Practice. *Bioethics* **13** (3-4), 314-326.

Commentary on Case Study #3

Steven N. Goodman, M.D., Ph.D., Johns Hopkins University

Dr. Goodman is an associate professor of oncology, pediatrics, epidemiology, and biostatistics at Johns Hopkins Medical School and at the John Hopkins Bloomberg School of Public Health. He is also a faculty member at the Johns Hopkins P.D. Berman Bioethics Institute and the Johns Hopkins Center for Clinical Trials.

The back-story of this trial is not just the design of the trial but that it represents the investment of decades of relationships that Dr. Weinstein developed with investigators all across the country. A certain degree of trust had to develop before this trial could be developed, and it represents a significant achievement. I run a project trying to find the most important RCTs in all of medicine, and we will certainly be looking at this trial when it is over.

There exists much confusion about how to navigate the continuum between explanatory trials and pragmatic trials. The SPORT trial affords a good opportunity to talk about some of those things. The kind of confusion that has motivated this conference is often due to disagreement or lack of clarity about foundational issues. In this situation, those issues concern when group summaries are meaningful and the foundations of causal inference.

I want to share a story from a consultation I had a week ago with an investigator at Hopkins; I came to him with the bad news that his trial for a drug was completely negative. I presented the data to him, and he said, “Why are you saying this was

unsuccessful? Forty percent of the treatment group responded." "But," I said, "40 percent of the placebo group had the same response." He said, "Well, the placebo worked for them." He was serious, and proposed a follow-up design that would confirm his prediction by having a run-in period with only placebo, and then removing the placebo responders. This second trial would be clearer because only the people who responded to treatment would be left.

What that incident illustrates is how hard it is for many physicians to think probabilistically, instead of in terms of deterministic mechanisms; if the patient improves while on a particular therapy, there must be some mechanism in each individual case that explains the response. Looking at numbers over groups to understand what is going on (or not going on) in a particular patient makes no sense.

These issues have been discussed in many forms for hundreds of years. Some people have traced these discussions to famous debates in the French Academy of Sciences, where they were talking about the prognostic numbers for urologic surgery. Claude Bernard, a French physiologist, was essential in moving medicine into the scientific era. This is what he said about probability:

> "A great surgeon performs operations for stone; later he makes a statistical summary and concludes from these statistics that the mortality law for this operation is 2 out of 5. I say that this ratio means literally nothing scientifically, gives us no certainty in performing the next operation. What really should be done instead of gathering facts empirically is to study them more accurately, each in its special determinism. We must study the causes of death with great care and try to discover in them the cause of mortal accidents so as to master the cause and avoid the accidents. Empiricism precedes science [counting knowledge goes before actual scientific understanding]; never have statistics taught anything, and never can they teach anything about the nature of phenomena. Statistics teach absolutely nothing about the mode of action of medicine nor the mechanism of cure."

I sometimes present this as an example of pre-modern, old-time thinking, but on some level I agree with it. What Bernard is talking about is the importance of understanding mechanism in order for numbers to have any meaning. He is talking about some of the same things we are encountering in this trial and the reasons that surgeons have been resistant to change on the basis of previous trials in this area.

Let me illustrate the issue with the example of insurance. My insurance company has a very good read on my prospects for living through the next 20 years. You can take my premium and put it over the payout and you would know the upper bound. If I went to another insurance company, you would find another highly accurate – but different – number. They bet millions of dollars that these numbers are right, and they don't lose money. But my doctor doesn't use those numbers. This is an example of how empiricism can get things "right" but gives us little guidance about the causes of risk in a particular case.

Let us now review the probabilistic definition of “cause,” which is not the kind of cause that Bernard and many physicians conceive of. For an individual, if the probability of disease or some outcome, if you have a certain factor, is greater than the probability of the same disease when you do not have that factor (which could be treatment), all other things being equal, then the factor is a cause of the disease. This definition allows you to have that outcome when you do not have the factor, and not have the outcome when the factor is there, making this a funny definition of cause, but that is what we must live with in today’s multifactorial causation world.

The problem with this definition is that we cannot identify cause in any particular case. Instead, we use group-based numbers to define it. We get a group of people who have a certain factor (for example, people who are treated in a certain way) and we want to see what their outcome would be if they did not have that factor. You want to see Patient A having gotten spine surgery and then roll the clock back and see what happens to that same person not getting spine surgery. Of course we cannot do that, so we get another group of patients who we say are “like” A, B, and C and we take their average, which is the basis for causal inference.

What are the bases for applying group averages to an individual in that group? What are the bases for having those numbers be meaningful? First is that there is no known causally relevant factor that makes any individual different from the group in some meaningful way. My insurance company is always right for the group but not necessarily for an individual within it. What does it mean to be right for the individual? You would have to get a thousand of that exact individual, which is not possible. What would be the closest thing to that? We would have to agree on a group of people who really were “like” the individual. But, what does “like” mean? We would have to agree on what is causally relevant; if we could agree on that, then that group would be the group whose number would be cited for the individual, and that would be meaningful.

What are the bases for applying RCT results to the individual? The basis is that there is no presumed interaction between treatment and patient characteristics or circumstances; that is, I want to know if there is anything about me that would modify the treatment effect (called “an interaction” in biostatistical jargon). The basis for the above inference is a presumption that all treatment effects occur via a common mechanism – the basis for the application to an individual is thinking that we understand the causal relationships within the group, i.e. the mechanism of action. In the absence of actual understanding, we presume, in the absence of evidence to the contrary, that all those in the group share the mechanism.

With that as background, let’s talk about surgery and surgeons. What makes surgery different than many other disciplines? It really is different, because surgeons must have confidence that, when they are in the operating room and they cut into the bowel and reconnect it a particular way, that this maneuver will have a reasonably predictable effect. If the patient is bleeding and a surgeon has to make an emergency decision, s/he must be confident of understanding enough about anatomy and physiology to make these decisions. That is the entire foundation for surgery. In many ways, it is different than the basis for predicting outcomes from drug treatments. Surgeons have a

strong *a priori* belief in their ability to predict what will happen if they make various interventions on the basis of mechanistic reasoning and prior experience. Therefore, each case can be unique, and "knowledge" is still claimed. The sense that surgeons know that a particular procedure will have a certain outcome is rooted deeply in the discipline that enables them to have confidence to act, both inside and outside the operating room.

Now we are prepared to discuss the paradox of "usual care." If practice heterogeneity is driven by a strong enough belief in customized care (as contrasted with differing beliefs about what is good for everyone) – if you believe strongly that it is an N-of-1 situation every time – then no group RCT result will disturb that belief. The ability to reason in the individual case in a deterministic way and risk predictions over groups are fundamentally incommensurable. A tension exists. Obviously, neither has complete claim to the truth, so we have to figure out how to negotiate this tension.

With this as background, I want to offer a few comments about the lack of difference between explanatory and pragmatic trials. Explanatory trials involve a proof of mechanism. They involve tight restriction on patient and treatment to minimize the possibility of interaction so that what you are getting is a causally homogenous group, but they still have a substantial degree of clinical and biologic heterogeneity. If you compare this to many laboratory experiments, they are a complete mess. The lab scientists often look with derision on all of clinical science because of the heterogeneity. What we call "tight explanatory trials" can look to the laboratory scientists like a complete biologic soup. When we draw the line between explanatory and pragmatic trials, we are drawing the "heterogeneity" line in a different place. Other people draw the line between RCTs and observational studies. It is completely arbitrary. You can see heterogeneity wherever you want to see it. The question is, when does it veer into the realm of not being interpretable?

What are "pragmatic" trials? The treatment mechanism is presumed the same as in the explanatory trial; that is the foundation for the trial. However, we expand the scope of potential interacting factors: patient characteristics, comorbidities, treatment characteristics, co-treatments, and setting characteristics. All these factors introduce the possibility that the effect of the treatment decision may be modified. These are often designed to inform us about the "interacting" factors. If there are none (if the pragmatic trial comes out with the same result as the "explanatory" trial), we say that the explanatory result is generalizable– that none of the things that occur in practice appreciably change the effect. Or, pragmatic trials can allow exploration of patient treatment subsets so we can understand why it is that different results occurred compared to the explanatory trial.

What about choice issues in the usual care arm? If you allow choice in the usual care arm, randomization no longer protects the subgroup inferences. This is not exactly equivalent to the large, simple trials in the sense that the interactions cannot be explored because the treatment is confounded with patient or disease characteristics. People who are less severe might choose one form of treatment and people who are

more severe might choose another, so it is more difficult to look at the interaction between usual care and a single alternative treatment.

What are the requirements for useful “pragmatic” trials? (This is perhaps the most important conceptual point of this discussion.) The effect of all the varying factors would still tell a causally coherent story, through a common mechanism. This is part of having the numbers apply to the individual, so they are relevant to the patient and the doctor talking about treatment at the bedside. Changing tidal volume in the ARDSnet trial was exactly this sort of thing: when you know what happens at 12 ml/kg, you are informed somewhat about 10 ml/kg and about 8 ml/kg. This entire continuum tells a story. Compliance tells a kind of causally coherent story – you expect that when you take less of a treatment you will have less of an effect. Expanding patient eligibility criteria at a higher ejection fraction (in defibrillator trials) tells a causally coherent story. Different doses of a drug tells such a story.

A pragmatic trial becomes a “salad” trial – causally incoherent – when the patients or treatments are perceived as being characterized by a qualitatively different set of causal factors, i.e., different treatment mechanisms. We can develop numbers that apply to causally heterogeneous groups in the same way that my actuarial risk did, but it actually does not apply to any one person in the group because the group is made up of a mix of treatments with different mechanisms. For example, suppose you got a 20 percent survival on one treatment and a 20 percent survival on another treatment. Does that mean you can aggregate those two groups and say you have a more precise estimate? I now tell you that one treatment is chemotherapy for ovarian cancer and the other treatment is for ADHD. You would never mix those two because they are qualitatively different causal scenarios. Even though the spectrum is much closer in spine surgery, we encounter the same underlying issue.

What are the alternatives to pragmatic trials? We can have multiple explanatory trials, in the realm of meta-analysis, conducted with different protocols, different eligibility, different settings, and different treatments. The advantages are that the inferences are protected by randomization. The conclusions tend to be robust, but the disadvantage is that it is a very long process, very expensive, and unpredictable, and who knows what trials are going to be proposed. In fact, these trials may not mimic real-world issues, so this may not be a complete substitute for the pragmatic clinical trial.

The nonsurgical treatments in SPORT constitute an extremely long list, which could be perceived as highly qualitatively heterogeneous. Will the treating community believe that they represent a common mechanism so that we can talk about the effect on this group as “the effect of nonsurgical therapy”? Will they believe that it is a causally coherent spectrum of treatments? If not, will that disturb physician-surgeon beliefs, which are already strong? I don’t know the answer to that question. So, while I genuinely admire what has been accomplished in this trial, I await the results with some trepidation – fear that a tremendous amount of time and effort may have been expended to generate results that the treating community may not accept, that they may find impossible to interpret, or that they may find easy to ignore.

Panel Discussion of Case Study #3

Dennis O. Dixon, Ph.D., National Institute of Allergy and Infectious Diseases (NIAID)

> *Dr. Dixon is a mathematical statistician in the Biostatistics Research Branch of the NIAID. He serves as senior advisor for the design and analysis of Federally sponsored research on interventions to prevent or treat HIV and AIDS. He is vice chair of the NIAID IRB for intramural clinical research and chair or member of five DSMBs.*

I want to echo Dr. Goodman's initial comments about how interesting this SPORT trial is; there are many aspects on which we could focus. I am looking forward to seeing the outcomes of this study, including the investigation of differences between those who agreed to be randomized and those who did not. It is terrific that such a study has been conducted, because it will provide much interesting material. I would like to focus on one aspect of the trial – the notion of what makes up the usual care group (the nonsurgery group).

Conclusions from the results of the randomized trial: in the manuscript, a comment about the reason for not being too precise about the nonsurgical approach was so that the results would be more generalizable than those from other RCTs in which there is a narrowly constructed control group. That is true in some sense; on the other hand, the conclusion from this study would be that the surgery is either better than, worse than, or as good as approaches in which individuals and their doctors select from the list of options in Table 2 in the manuscript and the list of medications. This design will not provide a definitive or valid comparison between surgery and any other particular choice. That would be true even if there was a high concentration of choices on one alternative to surgery, because those decisions to use that alternative would not have been decided on the basis of randomization so there would be no way to address the selection factors and other confounding influences on that choice.

On the other hand, there is a possible problem with the heterogeneity in the usual care group, and this trial offers a good example of how we may be putting too much emphasis on that possibility. A minimum form of care was specified in the nonsurgery group, including physical therapy, counseling, home exercises, and analgesia. The specific choices among those might not have been the same, but the basic elements of the nonsurgery approach were specified. On top of that, there were lots of other options that patients and their doctors could choose. It is worth asking whether the basic specification in the nonsurgery group would account for most of the observed efficacy, so that once those are specified there might not be much heterogeneity among the remaining choices. What matters is not exclusively the fact that the different options are described differently; it is the effects of those different options. In a usual care arm of a study, if there is some attempt to ensure that the options are not grossly inferior, then there is reason to anticipate that the outcomes might not be so heterogeneous.

For example, in Dr. Swanson's presentation yesterday, it was not surprising that he reported a lack of significant variation in the two groups. It may be because there was little heterogeneity in the effects of the choices in the usual care arm.

Alex John London, Ph.D., Carnegie Mellon University

> *Dr. London is an associate professor of philosophy and an executive member of the Center for Advancement of Applied Ethics at the Carnegie Mellon University. Dr. London is co-editor of Ethical Issues in Modern Medicine, 6th Edition. His research focuses primarily on foundational issues in research ethics, including issues in clinical trial design and the conduct of international research.*

Putting myself in the position of someone who might be asked to enroll in this sort of trial, I found the feature of this trial that was most disconcerting was the underlying portrait of the state of the nonmedical treatments for back pain. I only have several remarks to make and they will be brief.

First, I want to echo what Dr. Weijer said yesterday about clinical equipoise. We tend to focus on the importance of uncertainty for initiating a clinical trial, but equipoise has an equally important role to play in influencing the design of the trial. That is, clinical trials should be designed with the aim of generating information that will disturb, alleviate, or remedy uncertainty that exists within the clinical community about how best to treat a particular medical condition. In order to do this, trials must focus on the issues that are salient for clinical decisionmaking. In other words, they must focus on the endpoints that drive clinical decisionmaking in order to maximize the likelihood that the information they generate will produce the appropriate change in treatment behavior.

This last consideration reveals an important role for the preferences of patients in the design of clinical trials. If you study endpoints that are orthogonal to patient preference, then the results of such a study may not have a significant impact on clinical practice to the extent that clinical practice is itself responsive to the needs and concerns of patients. So the decision about what endpoints to study should take patient preferences into account as much as possible in order to ensure that information generated by the study will address the various factors that drive clinical practice surrounding the treatment of the medical condition in question.

Having said this, however, the standard that a single trial should change clinical practice is unreasonable. In the philosophy of science, for example, the idea of the single, paradigm-changing critical experiment has been shown to be something of an exaggeration. As a result, it is probably more appropriate to ascertain the relevance of a particular clinical trial to the considerations that actually influence clinical practice, and then to evaluate the value of that trial as part of a sequence of investigations that aim to achieve the goal of improving clinical practice.

I want to emphasize, therefore, that I share Dr. Goodman's and Dr. Dixon's concerns about whether it might have been possible to regularize the interventions administered to participants in the control group. If that was not possible, however, then we are left to consider how SPORT might contribute to changing clinical practice.

Here, I think it may help to view SPORT as one in a possible sequence of clinical trials. Given the degree of variation there appears to be in the treatment of back pain, it would be significant, for example, if SPORT were able to provide greater clarity about (a) how commonly used various treatment options are, (b) what the potential outcomes are under these various treatment scenarios, and (c) perhaps help to clarify what some of the underlying factors are that determine the utilities of patients over these outcomes. If SPORT can help in this regard, then it will be providing a beneficial contribution. In particular, because what physicians say and what they do can differ dramatically, SPORT might be able to provide a measure of the actual diversity of treatments as a first step toward regularizing or improving the medical treatment of back pain. If it turned out that the degree of variation in usual care in this trial was not as high as it might be, given the expansiveness of the menu from which people could choose, then that would also be an interesting discovery that would affect what we can infer from the trial. Although I share my colleagues' worries, if we view SPORT as the first trial in a potential series of trials (focusing on the ability of this trial to generate more concrete hypotheses for future investigation), then SPORT may have considerable merits despite some of the limitations on what can be directly inferred from this trial given the heterogeneity of the control group.

Dr. Levine: It is interesting and important that Dr. London called attention to the fact that expecting one clinical trial to change clinical practice is probably unreasonable. The late Tom Chalmers often bemoaned the fact that even multiple, good, decisive clinical trials seemed to have little effect. I particularly recall his having written several times about the demonstration that antacids had no effect on gastric ulcers. He tracked the practice patterns after these reports and found they had no impact whatsoever on the prescription of antacids.

Jon D. Lurie, M.D., M.S., Dartmouth Medical School

Dr. Lurie is an associate professor of medicine and of community and family medicine at the Dartmouth Medical School. Dr. Lurie currently practices hospital and general internal medicine at the Dartmouth-Hitchcock Medical Center. He is also a physician investigator on the SPORT trial and a member of the Dartmouth Atlas of Healthcare working group.

The fundamental point on which I agree with Dr. Goodman is that, when trying to make decisions about the appropriate control group and the appropriate study design, those questions hinge fundamentally not on general distinctions between efficacy and effectiveness or explanatory and pragmatic, but on the scientific question at hand and all the gory details associated with that. I talk to medical students when they come on the wards, and they are given classes on how to take a medical history. I tell them that it is nonsense; you learn to take a medical history by learning medicine and knowing what is going on with your patients. That is what tells you what to ask, not a card you carry in your pocket that says what you are supposed to ask. The same is true when trying to pick out control groups – it depends on the specific question to be studied, not whether you are formulating this trial to answer an efficacy or effectiveness question.

SPORT has three characteristics that pose fundamental challenges:

- It is trying to study surgery. There are heterogeneities associated with it that are different than trying to give someone a pill. You can make a same-colored pill that has something else in it, and it is presumed that someone in Iowa gives that pill in exactly the same way as someone in Florida. In the case of surgery, no surgeon thinks they perform surgery exactly the same way as someone else; they "know" that they perform it better!

- It is trying to study established treatments, to which the majority of people think they already know the answer because these treatments have been used for a while. It is easier when using a novel compound or treatment because people are less entrenched.

- It is a longterm study in free-range humans. ARDSnet had to struggle about their control group because their control group was going to get exactly what they decided they were going to get. We struggled over what our control group was, but we also knew that whatever we decided, that for the 5, 6, or 7 years of this trial in free-range humans, they would get whatever they wanted. If they were not better, they would choose to get acupuncture or glucosamine or Doan's pills – whatever people do with back pain that does not get better. The choice was not between having a homogeneous and a heterogeneous control group; it was between having an overtly heterogeneous and a covertly heterogeneous control group.

We could have said we were going to randomize people to X and would not "peek behind the curtain" as to everything else, and we could say we then have a nice homogeneous question to answer. Or we can say this is really what is happening and yes, it has challenges for interpretation. I would caution about the use of "uninterpretable," which has been bandied about this room. As Dr. Wittes pointed out, there are some people who say if there is crossover in an RCT, it is uninterpretable; it is not uninterpretable but rather is prone to potential confounding or it could be challenged. If you lose the strong protection of randomization, you have an observational comparison, which is not uninterpretable; it is merely prone to potential confounding. That is not what we would prefer, but we should avoid having perfection be the enemy of the good; as Dr. Wittes said yesterday, "better an approximate answer to the right question than a precise answer to the wrong one."

Discussion

Dr. Levine: Dr. Weinstein, how many surgeons refused to cooperate? In my experience on DSMBs, if it is a comparison of surgery to something else, at least one surgeon will claim that equipoise does not exist. This was the case in the Collaborative Ocular Melanoma Study and the National Emphysema Treatment Trial. I mention those two only because people connected with those DSMBs are in the room, but there have been others. I have not seen this occur when a study is drug treatment versus drug treatment. We have heard a lot from several of our panelists about how each surgeon thinks he or she is the best. How much of a problem was this in the SPORT trial?

Dr. Weinstein: Dr. Goodman's comments about surgery being different, and Dr. Lurie's echoing of that make a critically important point. Not many of you would want me to be your surgeon if I walked in the room and said, "I do not know what I am doing and do not know the answer but we can do this on Thursday." You probably would not sign up! It does present a unique challenge. Most surgical trials in spine surgery are comparing treatment A versus treatment B, both of which are surgical. There is a lot of compliance with that sort of approach, although the surgeon still holds the belief that their treatment is the better one. I spent 25 years working in this area and developing relationships with colleagues who were willing to suggest that they had equipoise and were willing to participate with the protocol as designed. Not everybody did, so not every Center had full cooperation and not every surgeon in every Center wanted to participate. However, it was clear in the beginning that, if you were going to participate, you needed to participate with the protocol despite its heterogeneity. We are proud of the success.

We realized the limitations, but we were crossing a mountain and it was not easy to get up the hill, and I still feel the rocks coming back down on me. We have a long way to go, but the inertia we have overcome by doing this is incredibly important to the public and to my patients. I do not regret the limitations nor do I apologize for them. We could always do things better, but I am struck by the cancer trials or any other trials in which I have been involved in which people seek other treatments. That fact is never talked about in those trials nor is it recorded anywhere. I would question the purity of any trial. We decided to be very open about what people were potentially going to achieve and receive. Dr. London's point is probably right – we will come up with some common, nonoperative treatments that people use that will be interesting, although I will still question their efficacy. We will move forward. It was a difficult issue of surgical willingness to participate, because it challenges what they believe in and what they do every day, which is a difficult situation.

Dr. Atkins: It seemed like much of the discussion was around the heterogeneity of the control group, but none of the surgeons would believe that any of those interventions was particular effective. The important part of being in a control group was that they did not get surgery. What exactly they got would not be particularly important, since the equipoise we are trying to disturb is the surgeons' equipoise. We are coming into this trial with an assumption that surgery is being over used; the question is whether we can define a group for whom surgery does not offer a benefit over no-surgery. If the outcomes are equal, the surgeon is not likely to say that is because they got such good physical therapy; they will probably say they operated on the wrong people. If 20 percent of both groups get better, surgeons will say they can pick out those 20 percent who will do well with surgery. Maybe the problems in this trial are more about surgical trials than about usual care groups. What would it take to change surgical practice? What kind of outcome would change what they do?

Dr. Weinstein: These are all the discussions we have had with my surgical colleagues and others. No one trial changes practice, even when it is really good. As we talked about yesterday, even though we know beta blockers or aspirin work, patients still do not get them – and they are simple. As a practicing surgeon, I can do a better job informing my patients with real information. I consider all of my studies not as studies

per se but as knowledge that I can impart in my practice every day to be a better clinician for my patients. I hope I can provide better information to my surgical colleagues that they can then impart to their patients; I cannot change their beliefs.

Dr. Djulbegovic: I want to come back to the tension between individual versus group, in terms of making inferences, generally, but also as in the SPORT trial, trials build in decision-making mechanisms for which treatments to choose. This kind of trial is not only about the issue of scientific inferences or ethics, but also about the issue of when it is rational to choose randomization as a decision-making tool to decide between surgery and a competing treatment. The most rational way to use randomization as a decision-making tool is when you are expecting a benefit-risk ratio of about 50 percent. We do not have a track record of a particular intervention we want to test before we do the trial, so we use that global assessment of equipoise or uncertainty as a surrogate for nebulous or poorly defined uncertainties. I would bet that the 33 percent of your patients who agreed to be randomized used randomization as a way to resolve the tension. They just did not know how to decide which treatment to select. They used randomization as a tool to decide between A versus B, and you get an almost-perfect split later in the observation arm. It is similar to some of the other big trials done in England, in terms of using uncertainty as a decision-making tool and also as an eligibility criterion.

Dr. Weinstein: I was surprised that so few people chose surgery; my colleagues thought that everyone would have chosen surgery given how much pain they were in. There was quite a range at the different sites as to who actually got surgery. That goes back to Jack's maps – the region predicts what will happen in the region. Jack and I talked about surgical signatures; that is an interesting phenomenon that you see in different regions of the country. My surgical colleagues are very anxious about this study. I do not care what the results are; I care that I get good information to impart to my patients.

Q: I would like to focus away from the information necessary to change positions and opinions. Surgeons are not the only humans with strong opinions! Those of us in pulmonary critical care medicine, ARDS Network, and other projects of which I am a member have extremely strong opinions. So what I am about to say should not be viewed as something achieved in a group that is different enough from surgical opinion-makers to be in a different world.

I would like to focus on internal validity – information you want to inform discussions with your patients should be based on internally valid results. We should ask a few things about science. Science is different from other human intellectual activities primarily because we experiment. One of the prime requisites in science, before the incorporation of new information or results into the body of knowledge of the domain (e.g., textbooks), is replicability. A chemist expects another laboratory to be able to replicate her results when a new observation is made. The basis for replicability is an adequately explicit method; if you do not have a method, you cannot replicate the experiment. I want to address, from the internally valid perspective, the methods that

you used. I admire enormously what you did and I thought the paper was wonderfully written.

This is an open trial. Our model for clinical trials (the blinded randomized trial) was based since the 1940s on drug trials. In open trials it becomes more critical to deal with confounders, particularly those that appear after randomization, having not been corrected by the randomization. Confounders can introduce enormous systematic bias that will not protect against a whole host of issues. I want to make note of the difference with which you approached the standards with regard to followup visits (there were rules for followups), the standards in surgical therapy (there were rules here too), and for teaching (the video viewed by all patients); however, with regard to medical therapy you elected not to list the sequence in which the 50+ things might be chosen. That is a deficiency that I hope might be considered in the future.

Q: A former mentor of mine said that the need of the healer is to believe and the need of the scientist is to doubt; I believe that tension is inherent. My question is a procedural question having to do with specifying what "usual care" is. One way to do that is to do a survey in advance to specify what is usual care, including what people do on their own and not only what is prescribed. The second approach is to document clearly what occurs in everybody – what kind of things they seek out, what kind of treatments they actually partake of. If there is a systematic difference between the two groups in the nonmedical therapies received, that could account for a treatment difference that could be falsely attributed to surgery.

Dr. Weinstein: Agreed.

Q: Dr. Goodman, I was concerned about your definition of probabilistic causation, which I believe in most of the world is called correlation and positive correlations. I realize that you realize the difference, but in the medical community, surgeons believe deterministically whereas psychiatrists believe probabilistically. It is critical to be careful how we use our words or people will misuse them in the future, as they have with usual care and some of the other specific areas. It is vital to be precise about which definitions are being used.

Dr. Goodman: I agree. And I agree that surgeons as well as other physicians often think deterministically. Most of us think that if we repeated the same procedure in any particular person, whether surgery or even medicine, we would see the same result again. We do not think there is an implicit risk. But the definition I posed was not correlation; it was probabilistic causation. It said that with the same individual and exactly the same counterfactual circumstances except for one thing changed, the outcome would be different. That actually is the definition of causation and not correlation.

Q: I have a question about the inclusion of degenerative spondylolisthesis people. How could they choose surgical or nonsurgical procedures? It is a degenerative condition that continues and cannot be stopped by taking drugs or by exercise. On page 1370 of the article, under "treatment crossover," you write, "although rates of crossover from

surgical to nonsurgical treatment have been substantial in prior studies." How could a person who has had a laminectomy go from that state to a nonsurgical state? Four years ago I had DS to the point at which my spinal chord was almost severed and I had to have surgery. I had two surgeons, one on each side, both trained at Wake Forest University – there was no choice. I could not have been randomized.

Dr. Weinstein: They may have been assigned surgery and decided not to have it. On the question of degenerative spondylolisthesis, you are correct. It is a process of aging, some people might say, more common in females at about 4 to 1. The issue there is, which is better. You may not change the degenerative or aging process, but does the surgical versus nonsurgical treatment effect a better outcome over time?

Q: I am a biostatistician. We have to remember that clinical trials are the art of the possible. In each setting, only a certain range of things are possible and then you have to figure out, in the range of the possible, whether you can answer a question that is worth answering. This was a great example because it was so different than the ARDS trial. In this trial, anything but usual care was impossible, given the long-term followup that was necessary and the fact they are free-range humans. You were forced into a usual care arm, but are the problems with usual care so much that it's not worth doing? I agree that it is worth doing.

The ARDS Network is almost in the complete opposite situation. In a good proportion of our units, our investigators are it – they run the unit, they attend to the patients, etc. If they designed and are conducting a protocol and have trained all of their staff on this protocol, it is rather difficult for them to suddenly do usual care. So in our setting the opposite was true and usual care was next to impossible. Remember this fact when making guidelines for clinical trials.

Dr. Levine: I would like to package this comment with the last comment, because we may have missed an opportunity. The woman who made the comment that there was "no alternative; had to have surgery" – is it compatible with the whole notion of usual care that the competent practitioner would have recognized that and advised surgery, so she would end up with surgery anyway? Or does this in fact undermine our confidence in a usual care arm?

Dr. Weinstein: What we've seen is that a lot of people with DS were willing to randomize – about 300 of them, across 11 States. We enrolled in the study about 86 percent of the patients who were eligible. At an individual level, people are different; I do not want to comment on an individual case without knowing the facts. Usual care around surgery and the indications for surgery are not something we discussed, but that is why we made very specific inclusion criteria, for which that person may not have been eligible.

Dr. Lurie: It is important to point out that there were exclusion criteria, such as progressive neurological deficit, where it would not be competent care to randomize that person to nonoperative treatment; that is one protection. If a person was randomized to nonoperative care and developed an unambiguous indication for surgery, they would

cross over and get surgery. That has its problems later in interpretation of the trial, but the patient's welfare is more important.

Dr. Levine: Let us say our speaker was eligible for randomization and was randomized to usual care (which is not the same as nonoperative care), and then she went to a competent practitioner. If she was in a situation in which there was no reasonable alternative to surgery, would not a competent practitioner advised her to have surgery? It would not be the case that, by randomizing to usual care, we are undermining the wellbeing of particular patients.

Q: I have a strong oncologic bias. One of the things I think should come out of this is some guidance about when the usual care group is too heterogeneous or when the evolution of the science is likely to cause that usual care group to change enough over the timespan of that trial. In those cases, you would have to consider using a usual care group very carefully.

Dr. Weinstein: We will learn a lot about the usual care group and what has been done. This will be the largest collection of data for this population on almost anything – surgical complications, operative procedures, usual care treatments, what they are actually getting, etc. We will learn a lot about this that will help in the future.

Dr. Goodman: It depends on how one positions the trial. If this was the last trial in the area, this might be seen as too heterogenous. Taking up Dr. London's point, if we see this as plowing both the methodological and ethical ground for further research and more refined questions, then what may be too much of a coherent inference for a single trial may be a very valuable first step in a long series of productive investigations.

Dr. Levine: I have to press the thought that whether or not this would be the last trial in the area. Suppose it turns out that surgical interventions are clearly superior to usual care. Now it is time to plan your next trial. Would you be surprised if the IRBs say that there is no question that surgery is always the best?

Dr. Goodman: When I was planning my talk, I tried to map out every possible scenario, including what would happen in the observational arm and what would happen in the randomized arm. We can paint particular extreme cases where we will come away with a relatively unambiguous answer. What we are likely to see is a moderate amount of heterogeneity in the control group and we will not know to what it should be attributed. We will also see different results in the observational group that might lead to some knotty inferential problems; for example, if we see better surgical outcomes in the observational surgical arm than in the randomized surgical arm. All I will say is that there are certain outcomes that could lead to somewhat unambiguous answers from this single trial. But the greater likelihood is that there will be complexities here that will not so much lead us to question the design of this trial but will lead to future trials.

Dr. Levine: Dr. Dixon, in your professional role, you have seen clinical trials where IRBs have said, "You should not do that because we already know the answer." Are you at liberty to discuss any of those?

Dr. Dixon: My experience is not so extensive so I am not sure I do have examples of that. The IRBs certainly will take account of the recent developments in the field, and would be reluctant permit a study to go forward unless there was some reason to think that the proposed intervention would be markedly different from what had already been studied and shown to be superior.

Dr. Wennberg: The fundamental background here is the concept that this is a close-call decision. It is a choice between competing outcomes. The choice should correspond to the patient's preferences, not the provider's preferences. The question we are after here is: what are the probability estimates we could get on the table? In the design of this trial, it is not quite what I wanted but it is close. I wanted to know whether people who actively choose have different outcomes from those who are randomized. That should be theoretically interesting and important. I am much more interested in the surgical arm of the open choice versus the surgical arm of the randomized trial. The trial I really wanted to do was to randomize to a randomized trial versus a preference trial, so we could be certain that, going in, everybody was the same. My hypothesis is that the utility gained would be greater among those who actively choose compared to those who are randomized. That is an important perspective that we need to get on the table. If we assume that we are looking for precision in terms of our estimates, then the circumstances under which those estimates are made, from the patient's point of view and psychology, are tremendously important.

To give you a sense of the context of this problem, there is an interesting study done in Canada in which all the clinical criteria (this was for knees, not for backs) were based on a population-based survey. According to the criteria of the clinicians, N people would be chosen as being eligible for surgery. However, when they were interviewed and offered surgery, only 15 percent of N chose it. That is important to understand.

What we are after here is not just a cross-sectional decision but also a longitudinal decision. People get worse or they get better, and there will be crossover. Crossover is rational and we need to know when that happens. We have a lot of different goals in this trial other than classic efficacy analysis. The usual care group is fine; we will get a lot of good information. Maybe we can do a better trial on the medical arm at a later date.

Roundtable Discussion: Development of Ethics and Scientific Principles To Guide Considerations of Usual Medical Care in Clinical Trial Design

Roundtable Chair:

Alan R. Fleischman, M.D.
National Institute of Child Health and Human Development, NIH

Roundtable Discussion: Development of Ethics and Scientific Principles To Guide Considerations of Usual Medical Care in Clinical Trial Design

Chair: Alan R. Fleischman, M.D., National Institute of Child Health and Human Development, NIH

Dr. Fleischman is clinical professor of pediatrics and clinical professor of epidemiology and population health at the Albert Einstein College of Medicine. Dr. Fleischman is chair of the Federal Advisory Committee as well as ethics advisor to the National Children's Study at the National Institute of Child Health and Human Development. He is also senior advisor to the New York Academy of Medicine, where he once was the senior vice president. Dr. Fleischman is a member of the New York State Governor's Task Force on Life and the Law and the Children's Subcommittee of the DHHS Secretary's Advisory Committee on Human Research Protections.

Our goal for this roundtable discussion is to have two 1.5-hour segments in which we look at a "points to consider" document. This issue has some contentious parts to it. Our hope is that we will be able to generate some levels of agreement and some places to define what we know and do not know, what we need to know further, and where there are honest disagreements.

Panelists introduced themselves, including their institutional affiliation and their professional identification regarding input to this process:

- Julie Zito, University of Maryland, Baltimore, pharmaco-epidemiologist (psychiatry and mental health)
- Zeke Emanuel, Department of Clinical Bioethics, NIH; oncologist and bioethicist
- Larry Friedman, retired, clinical trialist for many years at the NHLBI
- Steve Goodman, once a pediatrician; now a biostatistician, epidemiologist, and clinical trialist
- Bob Levine, ethicist
- Deborah Zarin, NIH, clinical trials registry; child psychologist, decision analyst, wrote practice guidelines
- Henry Silverman, University MD school of medicine; clinician, ethicist, ARDSnet investigator in spirit
- James Swanson, University of California, Irvine, developmental psychologist
- Anne-Marie Swart, medical epidemiologist for the Medical Research Council
- Taylor Thompson, clinician trialist at Mass. General Hospital and Harvard Medical School
- Charles Weijer, Department of Philosophy, University of Western Ontario, Canada; philosophy and medicine
- Jim Weinstein, Dartmouth; orthopedic surgeon
- John Wennberg, Dartmouth; epidemiologist
- Janet Wittes, statistician, Statistics Collaborative
- Dennis Dixon, NIAID, biostatistician
- Brian Haynes, clinical epidemiologist and clinician, McMaster University; "informologist"

Ground rules for this activity:

- Trials involve prospective assignment to a treatment or intervention group – not just clinical trials but also health services intervention groups or other kinds of interventions.
- Trials need to be relevant to clinical practice.
- Trials need to have two or more comparison arms.
- Trials must have health-related outcomes.
- This discussion will not include observational studies or the international context (which would add another layer of complexity that will be dealt with at a different meeting).

The strawman document in your packet was generated by staff and professional input, and it is not going to be wordsmithed but will be added to during this activity. We should try to agree on some definitions. We talked about usual care, standard care, and standard of care. For the sake of our conversations, I would like to get agreement that the concept of usual care and standard care are the same. Standard of care elevates to a care pattern that is generally accepted, whether or not it is evidence based and whether or not the clinicians use it.

Dr. Zarin: The concept that came up yesterday, "competent care," sounded appealing to me. When you talk about usual or standard care, are you thinking about a range? In my mind, for many conditions, usual care includes a range within which we will all agree we would like to think people are operating.

Dr. Fleischman: Community care is the same as usual or standard care – it is a broad range of "anything goes." It is the "wild type."

Dr. Levine: We will not be using the term "standard of care" in the points-to-consider document or in the rest of this discussion, so it is not necessary to define it. However, I usually see it as a legal term – the established minimal standard below which you may be found negligently liable. It has also been applied to evaluating professional behavior outside the field of medicine.

Dr. Thompson: "Standard care" is a confusing term; we should decide not to use it.

Dr. Fleischman: Usual care will be the wide range of what happens in the community (competent or not, evidence based or not).

Dr. Zarin: Usual care is a package of services that the community of clinicians would expect from X percent of well-regarded clinicians – what we would all consider basic care.

Dr. Silverman: I would describe it as unrestricted practices of healthcare providers. What about protocolized usual care? In the SPORT study, we had some commentary

on the defined minimum of care in the usual group and one could claim that that was more akin to protocolized usual care, which bears some consideration in design.

Dr. Fleischman: Protocolized usual care would mean impacting on what happens naturally in the environment to try to control or take a subset of it, and leave usual care to be broader.

Dr. Swanson: Another definition is "the record of care that is provided in that group." In the MTA study it was a community comparison, and we grappled a long time with how to measure what was obtained. We came up with instruments to measure what came up.

Dr. Fleischman: If usual care is part of research, we will have to have methodologies to figure it out.

Dr. Weijer: I have heard a mixture of descriptive and normative language around usual care, and I encourage us to keep those clear. Usual care could be understood as a descriptive term – the care that is provided in a community. We keep that separate from notions like competent care or standard of care, which would be a normative assessment of how much of that scope of practice actually delivered ought to be delivered.

Dr. Friedman: The first time I came across the term "usual care" was in some of the early NHLBI clinical trials, in which the comparative group was labeled, literally, usual care. The assumption was that sending people back into the community to obtain care would not be as good as what could be provided in the clinic, and therefore allowing a true comparison. In some sense, that definition emphasizes that usual care is not necessarily good, best, or even competent.

Dr. Zarin: I am still troubled. What if you are doing an oncology study and your usual care arm includes the one physician in your State who believes in using Laetrile?

Dr. Fleischman: Since this is descriptive, that person is in usual care. We might argue that it is not standard of care or it is not competent care or it is on the far extreme of care, but it is in the community. For the descriptive piece, it is in. Whether we want it in as we move forward into describing how we use usual care in the research context is the next question. When can or should usual care be part of the research context of an intervention or clinical trial? "Can" and "should" are both important parts of that.

Dr. Goodman: It depends critically on whether the reason for using it is epistemic or logistical, which corresponds to normative versus descriptive characterization. If we are using usual care, as in the SPORT trial, because it is one of the only ways to implement the trial, that is one category of research. We cannot do the trial if we do not incorporate usual care. If the reason usual care is being invoked is that it is thought to be optimal or desirable or to have epistemic force (as in the ARDSnet trial where it was put forward that we were evolving to this new standard because it was a better standard), then that has ethical implications and is a distinctly different issue. We have

to separate those two situations about why we are where we are, which is spoken to in the background document to some extent.

Dr. Dixon: The key is whether or not there is a consensus in the relevant community as to the preferred treatment and what should be used as a comparator for evaluating an innovation. If there is no consensus, then it makes sense to use a usual care arm in which various options are made available within the trial.

Dr. Silverman: Maybe the first question is, should it be used? And then, can it be used? In terms of should, it depends on the research question being asked. Do you want to know what treatment should replace usual care? Then could you design a trial within the context of the heterogeneity of usual care practices, based on the type of disease you are studying and some other contextual issues? Then you would ask and answer the question of whether you can include the usual care group – considering feasibility in terms of clinical values, validity, efficiency, and patient safety.

Dr. Wittes: If we are talking about the kind of usual care that is nonprotocolized, if we believe all those are equally ineffective or maybe equally effective, then having that *gemisch* of usual care is fine and desirable and probably useful.

Dr. Levine: Regarding the "should" part of the question, usual care as a comparator should be considered only in high-stakes clinical trials. We do not need to engage in the expense and complications of a usual care arm if we are looking at, for instance, a new analgesic to treat headaches; it is not worth it. High stakes, in the sense of what we are trying to deal with, is something that can result in either death or disability or both.

The second attribute is that using a usual care arm should presuppose a reasonable null hypothesis. It has to be plausible to postulate that the intervention comparing to usual care could be no better than or could be worse than the usual care arm.

Dr. Haynes: The way to approach this would be when prior trials have shown that the intervention has shown sufficient promise in terms of benefits, harms, or costs to warrant consideration for widespread introduction of the usual care. It would be the basis for considering whether you should put the effort into trying to make full-scale implementation of practice.

Dr. Emanuel: Look at the paragraph on the last page of the "points to consider" document we have in front of us. It says that the motivation usually is the need to establish the superiority, or non-inferiority, of either a new intervention or a competing existing intervention, to current medical practice through direct comparison in the setting of a randomized trial. I presume that the setting of a randomized trial already suggests that there is sufficient promise or uncertainty (promise of the new intervention and sufficient uncertainty as to which one is better) on any one of a number of criteria – effectiveness, side effects, or cost, all of which would be legitimate. I am not sure that I agree with Bob. There is some motivation for high stakes, but in a lot of cases you might want usual care because there is common practice and high cost, not necessarily

high stakes. Consider complementary and alternative medicine (CAM) interventions, where there is a lot of cost although probably not high stakes. Also high blood pressure questions, because cost is an issue rather than effectiveness or side effects.

Dr. Zito: In regard to Dr. Levine's suggestion about "high stakes," I wanted to define high stakes in terms of high utilization and relatively low evidence base.

Dr. Zarin: In coming up with this strawman draft, we have thought about ethical and scientific reasons and we could not totally separate them. But focusing on the scientific reasons, sometimes there is a question that could only be answered by a usual care arm. In the MTA study, for example, if there had not been a usual care arm, a reasonable person might have concluded that Ritalin works better than everything else. As a clinician I could then say, "I am already giving my patients Ritalin so I do not have to change anything." The presence of a usual care arm allowed us to say, "There is something else going on besides the chemical moiety being delivered to the patient that is important to understand." Usual care in this case is not the same as the medication arm. That was an important scientific question that could only be addressed that way. I guess this comes under Dr. Zito's revision of Dr. Levine's rule, which is high utilization and cost and a limited evidence base.

Dr. Swanson: The last discussion before this panel used the term "customized care" that was important to understand usual care. The opposite of that – the care provided in the study – typically is protocol-driven and is standard for everybody. So many physicians, not only in surgery but also child psychiatry, think they understand this particular patient and what is needed, and they go about it that way. The term "customized care" for the usual care arm was very revealing about what it might actually be in practice.

Dr. Levine: I was not clear in my definition of high stakes, but I would certainly include high cost in some circumstances as being high stakes. I disagree that calling for a usual care comparator presupposes a preexisting RCT showing efficacy. There are many situations, including the SPORT trial, in which a preexisting RCT could set up an obstacle to moving forward with a trial in which the comparator is usual care. Preliminary evidence obtained from such trials would not be regarded as preliminary by many people, including most IRB members, and they might see this as grounds for disapproving a new trial with a usual care comparator.

Dr. Thompson: I think the "can" question is: "it is easy, it has already been used, sure it can." It has been used in weaning trials: unrestricted, *de facto* usual care practices in clinical trial designs. We saw it in the SPORT trial. The "can" question is easier; the "should" question is more difficult. Maybe "should always" should be discussed – some have suggested that usual care "should always" be part of the clinical trial. The "should" question turns on the research question and what we just talked about – the nature of usual care. I agree with Dr. Levine that it is artificial to require a prior RCT to give credence to usual care as a necessary condition for its inclusion.

Dr. Fleischman: We have put that to rest.

Dr. Wittes: I want to disagree with the CAM example. It seems to me that when a piece of usual care is that people really believe in a particular treatment, then the comparison is a treatment versus the one they care about. If in "can" you let everybody do what they want, then you have the problem of heterogeneity of belief.

Dr. Haynes: The requirement for previous RCTs ought to apply where something is not currently available in practice. If something is in usual practice and there are concerns about whether it works or it is wasteful or harmful, then the comparator should also be a usual care trial without the necessity for a prior RCT. These are not mutually exclusive at all; they may be complementary.

Dr. Goodman: It critically depends on what is causing the heterogeneity seen in usual care. Is it haphazard? Is it lack of understanding of the evidence base? Is it a lack of evidence base? Is it a specific belief that customized care is superior? Or is it a differing belief about standards that should apply to everyone? Each one of these will imply a different question and have different imperatives in terms of whether they should be included in the trial. It is absolutely critical to characterize the reasons for heterogeneity. This was made clear in the ARDSnet and in Dr. Wennberg's study.

Dr. Fleischman: I want to ask you this question: of all those different types of usual care, are they acceptable as part of clinical trials within certain contexts? Or are there some we ought to eliminate?

Audience: I would like some comment on the word "customized." The implication is that, somehow, usual care tailors decisionmaking to the patient's individual needs and protocols/studies do not. Protocols can take many forms. Most protocols and guidelines that are extant in healthcare delivery are not adequately explicit – they provide some guidance but not instructions about what to do. Some protocols can be adequately explicit and, at the same time, allow tailoring therapy to patients.

We have (as yet unpublished) data on the execution of physician intent in mechanical ventilation, in a multicenter trial using computerized protocols. It turns out the physicians are less satisfactory in carrying out what they want to achieve than a computerized protocol built on the rules captured from them through knowledge engineering.

Dr. Swanson: I want to come back to customized care and get to one additional point that might be related to training. This idea of belief might be variation in usual care. In my experience, you do what you are trained to do. You use stimulant medication because your mentor taught you how to do it. Trials ought to be intended to alter the scientific knowledge that is the basis for training. Customized care might be related to that topic and might be a target of a clinical trial for that reason.

Dr. Swart: Even in the United States, it is difficult to standardize all aspects of medical care, unless you have evidence to protocolize medical care that will make clinicians change regular practice for patients in the reference arm of a trial. If the intervention will work, it must work on the background of therapies that these patients would receive

anyway. The best idea I have heard in the last two days is from the SPORT trial – going through the process of defining which background interventions are important and collecting data on them. It makes you go through the process of identifying what treatment might be important. Then you need randomization of a sufficient number of patients to take care of the variability.

Dr. Weijer: It is the scientific question at hand that drives considerations of trial design. If that point is uncontentious, then I would suggest the following two points. There ought to be a clear and justified scientific question that drives the need for a usual care comparator. If that is true, then the following also ought to be true: that the usual care arm answers the scientific question in a way that a protocolized arm cannot.

Dr. Goodman: We have a real chicken-and-egg problem here. Do you start where the evidence is and then decide on the question, or vice versa? The clear characterization of the state of the evidence base is where you start. From that arises the question, because it cannot be characterized. The starting point is whether evidence is present and we have usual care practices as they are, or evidence is absent and we have usual practices as they are—which might be consistent in the absence of evidence and inconsistent in the presence of evidence. The presence or absence of evidence does not dictate the degree of heterogeneity of that arm; many unproven practices are remarkably widely homogenously followed. Once that is characterized, then the spectrum of reasonable questions can be outlined. How and whether the usual care arm is implemented is driven by that question. We have to start with a clear characterization of the evidence base.

Dr. Fleischman: I want to discuss the heterogeneity question from the scientific and statistical point of view. One might wish to answer a question – "Does this make sense from a scientific validity perspective?" – in the heterogenous world of usual care. I also want to look at the human subjects questions around usual care and whether there are differential risks based in usual care arms as compared to protocolized arms. Perhaps to start on the scientific question, I will ask Janet – I have a feeling that you have some skepticism about usual care heterogeneity.

Dr. Wittes: Actually, I am probably more enthusiastic about the heterogeneity than may have come out. With respect to statistical power, if we think that usual care is a mix of therapies (as in SPORT) and if we think there are differences in the effects of therapy and we worry that the experimental therapy will not hit a home run, then we have to worry about power. People will want to compare the new therapy to something that they can characterize.

As Dr. Britain pointed out yesterday in a question she asked me, there is a more fundamental problem – the problem of identifiability. You are comparing the active treatment arm to the usual care arm, and you do not know to whom to compare, among the patients in the usual care arm that have received all the various therapies. That is where one would worry about heterogeneity that might lead people to dismiss the results of the trial. That is why I brought up CAM, because people are committed to individual complementary medicines. If you are not worried about those and you want

an approximate answer to the question of whether this therapy is better than what is going on, on average, in the community, then I would not worry too much about power.

Dr. Wennberg: In the SPORT trial, our goal was to figure out the probability estimates that we should quote patients who are on the surgical side of the equation. A lot of the surgical outcomes are specific to surgery; they do not happen if you do not have surgery. Patients are interested in estimates of how long they have to wait before getting pain relief and things like that. The SPORT trial was an effort to make it possible for patients to make better decisions. There is complexity on the side of usual care, because there are many theories that have never been tested. I am not so worried about subgroup analysis within the usual care arm, because we are really interested in the global effects of usual care. It is a question of what is the threshold when people decide to undergo surgery, and how to convey information to them to help them make that choice. To know that, you have to know the probability aspects of surgical treatment better than we did.

Dr. Weinstein: I still go back to the pragmatic issue of what is happening in the real world. The reality is, even in our protocolized attempt (physical therapy, medication, etc.), the reality was that other treatments were going on. I do not see how ethically I could come forward with a protocol that says that A is better than B. The reality is that there is a lot more going on behind the curtain that we need to describe. It is an opportunity to understand that *gemisch* a little better. Our patients need to know what works and what does not. Any way we can get our hands around that, ethically, is really important. That does not mean that we should not presuppose another trial might be better designed, with tighter treatment arms, but I will still worry about what is happening around each side of those arms.

Dr. Friedman: I would like to extend that a little bit. If nothing in the *gemisch* of usual care makes any difference or it all has more or less the same effect, then there is no reason for the argument against it. The only issue is if we suspect there might be some differences. Just as we have talked about only certain types of people enroll in or volunteer for clinical trials, only certain types of practitioners conduct or get involved in clinical trials. The kinds of practitioners who do these trials are not a representative sample of practitioners in a community or in a country. So when we talk about usual care, we talk about usual care in those often-academic sites, which may or may not reflect usual care "out there." Even if there is homogeneity of effect among the practitioners in the study, there may or may not be homogeneity of effect more broadly. We are then trying to extrapolate the results. That is an added problem.

Audience: What if usual care is a dismal failure—rather than the case where it is pleasant and works?

Dr. Emanuel: If usual care really is as bad as you say, we have no obligation to provide it, ethically. You will not then be conducting a randomized trial including usual care unless you cannot get the community of practitioners to move off that for whatever reasons – reimbursement, inertia, etc. From the ethical perspective, if you really do have a treatment that fails 90 percent of the time, it is a real question whether you have

an obligation to provide that. Some of the cases where you see this are in emergency medicine and resuscitation out in the field.

Dr. Fleischman: The question is deeper than that; it is the question of the differential risks to subjects. If I sit on an IRB and I am worried not only about informing but about the therapeutic misconception, by including usual care in the research you are implying that whatever is going on out there in the world is acceptable. If you are the professor and you sit in an academic medical center, then the patient enrolling in the trial will respect your view of this difficult area. So when you (the researcher) come to me (an IRB member) and say you are going to let the "wild type" roam, I would be worried, even with your elegant and informative video, that subjects may be differentially placed at risk by your design. I have concerns about that.

Dr. Weinstein: There is risk on both sides of the equation, whether it is the surgical or the "wild-type," non-operative care that is not as wild as it sounds. The video looks at the probabilities of success and evidence for nonsteroidals, epidural injections, etc. Many treatments on the list are well described so that the patient is well informed, maybe better than their clinician who is giving them informed consent as a separate document from the trial. The knowledge base of any physician around all of the risk/benefit ratios and complications from drugs, injections, chiropractic treatment, etc. may not be at their fingertips; this helps to make that more uniformly available.

Dr. Levine: What if usual care is associated with a 95 percent death rate? If you already know that the thing you are testing will give you much better than a 5 percent success rate, then you are lacking one of the requirements for justifying a usual care comparator – the plausible null hypothesis does not exist. The field of oncology gives some insight – they try to use state-of-the-art chemotherapy as the comparator group with the hopeful expectation that the new regimen will prove superior. This is sometimes pressed to the extreme. I recall a clinical trial from 20 years ago studying a new therapy for renal cell carcinoma. Nothing did much good. They proposed to use a specific progesterone treatment in the control arm because there had been two case reports in which it appeared that the renal cell carcinoma got somewhat smaller. The IRB questioned them because this was not an established standard. But nothing else was available, so the statement was "we are using as our comparator the only thing that anyone has ever suggested might have some success." In extreme cases, I would say that even if the expected outcome in usual care is 95 percent fatal (or in this trial I am recalling it was 100 percent fatal), it might be quite appropriate if there is a plausible null hypothesis to use the usual care arm.

Dr. Emanuel: That oncology trial is a classic problem – oncologists cannot do placebos because they always have to do *something*; that turns out to be an expensive placebo. We ought to "fess up" to it. If you are not actually killing people with the intervention, it may be ethical.

Alan, I want to challenge to the way you objected to my characterization. We need to be very skeptical of data going into trials about the effectiveness of the latest new *gemisch* – whether it is surgery, radiation, or drug treatment. One of our tendencies is

always to overplay those things. Phase II data tends to be lopsided, for reasons that Dr. Friedman outlined, and we need to be somewhat skeptical. Clinical trials are littered with cases we are all sure would work until tested – and then it did not work or produced more fatalities. We need to be more skeptical about the new versus the usual care, where usual care has some plausibility or some data behind it. There are lots of cases in which usual care has no data; that is a separate issue. We should not just assume the new thing is better than usual care because usual care is marginal at best; the new thing may turn out to be better or it may be worse.

Dr. Atkins: It seems that the issue is that, given the greater heterogeneity one gets in using usual care as opposed to some defined care or some protocolized usual care, we need to have a reasonable belief that a heterogeneous arm will give us as good outcomes as the alternatives. In the SPORT trial, in which we are allowing people to seek any combination of usual care, we can postulate that since they are seeking pain relief they are probably going to do as well if we do not constrain that than if we try to develop some elaborate protocols to constrain it. We might exclude Laetrile from usual care. If we knew that usual care included 50 percent of people getting ineffective care, then incorporating that heterogeneity into the trial may not be particularly useful unless we are trying to answer a different question about what it takes to change practice.

Dr. Fleischman: But how do you do that? If we are going to decide what is acceptable usual care versus what is unacceptable usual care, how do we go about doing that from the investigator's perspective and then from the scientific review perspective and the IRB perspective? That will be the conversation this afternoon – about obligations and responsibilities.

Dr. Goodman: With regard to heterogeneity, you may have a different degree of evidence behind the null hypothesis with regard to certain comparators in the usual care arm versus others, depending on the heterogeneity. That is the crux of the problem. Here is a concrete example. Imagine that we randomized people in the usual care arm of the SPORT trial to meditation, massage, and chiropractice. It was not customized but still reflected some cross-section. Pretend we did a subgroup analysis comparing massage to surgery and got a p of 0.2. Then we compared ultraviolet light exposure and also got a p value of 0.2. If you put those together, they would get a p value for the combination of ultraviolet light and massage of 0.01. Is the evidence therefore compelling for the combination of ultraviolet light plus massage? Or do we have equivocal evidence about each individually? This is why I said that it is absolutely critical that we can believe that the mechanism underlying the two is the same. It is not enough to say that they might have equivalent efficacy; the evidence base is different. The epistemic status of each is different and the mechanism is different. This heterogeneity can be a real issue and there are some very knotty analytic problems that will arise if we just put them all together. I am seconding Dr. Levine's point that we have to look at the null hypothesis for every component of usual care.

Dr. Lurie: I am confused by your statement about risk in the usual care arm. From the point of view of a trialist, if usual care is really USUAL care, I cannot be putting them at increased risk. If it is really usual, how can there be increased risk? I, the investigator,

am getting out of the way of the care they would usually get, and anything I do to influence usual care has the potential to increase risk. But how can usual care increase risk over usual care?

Dr. Fleischman: By placing usual care into a human subjects research enterprise, you take certain responsibilities as the investigator for justifying putting those individuals in that arm. Your contribution to the bad care out there is different when you have been part of accepting that the subjects are part of that experience. As an IRB member, I can ask you whether the subjects of your research will believe that that is an acceptable level of care, or whether you will share with them your concerns about, for instance, Laetrile or other "wild-type" usual care. Or whether you will limit usual care to those who fit into some kind of acceptable evidence in outcome. You have to make that case. Subjects may fall into the misconception that you believe that the wild-type usual care is at least reasonable.

Dr. Lurie: Maybe there is a language problem. I would say there is a difference between putting people at increased risk and being responsible for what happens to them with the risk that they had before. Those are different. To call it "increased risk" is mistaken.

Dr. Emanuel: From the IRB perspective, when they are evaluating the risk, you are right: the question is the risk of care versus the risk of usual care. That is the IRB evaluation. Your responsibility as a researcher is a different issue.

Dr. Levine: When I said there must be a plausible null hypothesis, I did not mean a null hypothesis for each and every component of usual care. The null hypothesis could be put in these terms: that the intervention that we are evaluating will yield a better result than what patients can get "out there" on their own. We do not have to worry about evaluating each component of what they find out there. What people find out there in medical practice is something we do not control too much, but we rely on the FDA to rule out, for instance, Laetrile. We rely on malpractice litigation to rule out certain types of extremely incompetent medical care. We are saying that we think there is a null hypothesis at the outset that the intervention of interest is better or worse than what patients on their own can find in the competent-practicing community.

Audience: I would like to follow up on Pearl O'Roarke's door opener and I would respectfully remind Dr. Levine of the high error rate in medicine within the well-intended practice community. In the ARDSnet we have been concerned in designing trials about the result of a trial that might show that one comparison group might be better than what we posited at the outset of the trial. When we do comparisons in two groups, we try to plan for a good explanation of a negative trial and look for intermediate outcomes that indicate drug effects. What does the panel think if the pragmatic clinical trial deems usual care to be superior to the intervention? What would you explain to the medical community should be done, given the dynamic nature of usual care?

Dr. Fleischman: That is a distal question for later.

Audience: I think it is critical to always be careful to describe usual care as what physicians do versus what patients get. I reviewed a series of studies in which many of the patients got no care. To use an analogy, 10 percent of the patients get a wonderfully effective therapy and 90 percent get nothing due to lack of access, and this is called "usual care" in the studies. There is nothing wrong with a physician-centered approach but do not confuse that with studies that consign people in the community where they do not get access to care. That is a critical discussion – the difference between usual care in terms of what patients get and usual care in terms of what physicians do.

Audience: Do not forget the issue of statistical power. In the event of infinite numbers of patients, most people would like to include a usual care arm that includes as many choices as were available. The main study that brought this conference together was the inclusion of a usual care arm in an explanatory trial, where you do not have an infinite number of patients to make the usual care arm larger. The recommendations will be important for those of us who will not have the opportunity to keep this usual care arm quite as large as would be preferred.

Dr. Djulbegovic: Pushing the null hypothesis is ultimately a quantitative exercise. You really have to be well informed about your effect size.

Dr. Fleischman: I want to go toward a question of whether using the usual care approach has distinctions when we are evaluating several accepted treatments in common use versus a new intervention compared to usual care. There are a host of interesting treatment approaches in medical care for which the evidence base is not developed. It is common to do such things. Is there a distinction in thinking about this when evaluating two accepted treatments that are part of usual care versus usual care up against a new experimental treatment? Are there differences in the way we should think about usual care in those different scenarios?

Dr. Silverman: It depends on whether those two usual care interventions are representative of all usual care practices. If you just take A and B and you still have a lot of C left, then you may not be able to answer the question of whether A and B is better than usual care, which is represented by that large C. If both interventions capture or are representative of most of what happens in usual care, then you do not have a problem. In the second scenario, if you have a new intervention versus usual care, I do not see a problem there except for all the other issues we were talking about with usual care. That is one caveat I have with the two accepted interventions.

Dr. Dixon: You are saying there are two accepted approaches that are normally part of usual care and the interest is in comparing the two of them to each other?

Dr. Fleischman: Yes. There is some sketchy evidence of utility, doctors think they are useful and effective, and someone wants to do Usual Care A versus Usual Care B, with Dr. Silverman's caveat that there is not a bigger piece of C. If A and B are a major part of the action, is that different? The planning group was concerned about whether we

ought to think about usual care in those circumstances differently than we think about it when we take usual care against a new experimental intervention.

Dr. Dixon: I would say it is different to the extent that if you can organize a trial in which you randomly assign volunteers to A or B, then it is just a conventional randomized trial. If you are saying that you want to do a study in which volunteers and their doctors decide either A or B and then try to compare later, you would just be out of luck. Neither of those is similar to the model we have been discussing.

Dr. Zito: One distinction that could be made is if you are doing a comparison of a new medication versus usual care, you would have a very different knowledge of the safety profile of the two. Whereas when you are doing a number of established treatments against each other, then information about the safety profiles would be more or less similar.

Dr. Fleischman: We have dealt with this problem. I raised it because it seems within the usual care context to have some differences that people wanted to raise.

Audience: In the SPORT study, you are looking at one element of usual care, taking it out of the usual care and asking whether this element is better than all of the rest. This could dilute other possibly effective usual care elements. I work with a pain management specialist who does injections. That injection could be lumped into that "other usual care" in the SPORT trial. A pain specialist might send people who fail his therapy to a surgeon and vice versa. You have taken one important element out of that – the whole usual care picture. In other words, you have taken that one item – surgery – out of the mix and compared it to everything else.

Dr. Dixon: Just be careful which question you are trying to answer. If you want to get objective information as to whether the surgery has benefit, then the sort of study that was planned and discussed here is reasonable. It is much more difficult if you want to tease out and examine the experience of some patients who start with injections and then get the another treatment as salvage versus starting with the other and then getting injections as salvage. That will be complicated to interpret, but it is a matter of which specific question you try to answer.

Dr. Fleischman: I want to move us into the question of the distinction between wild-type usual care and protocolized usual care and competent care. Dr. Thompson, your trial and Dr. Weinstein's trial are so different. He is the "wild west" and you are inside the ICU. I wonder about the impact on usual care, for example the decision whether to protocolize it or to say "some are in and some are out." Any thoughts about how that fits in the research design?

Dr. Silverman: That gets to the heart of the issue. There is no one answer for that. The concept of competent care comes from the legal community, as does this notion of a standard of care – a normative description of a group of practice patterns that are considered competent by their peers. There are a number of layers of review that are required to set the bounds of competent care. It is a necessary subset of wild-type

usual care and there may be some ethical considerations to randomizing patients to groups that have wild-type usual care, with concerns about the quality of care on either extreme. The bottom line is that we have to recognize that that is a qualitative decision. If it is going to be made, that decision needs to be made in an open, transparent way so there is agreement about that.

Dr. Swanson: When you go through the informed consent about all the arms, you certainly inform the person and give educational materials about what each arm involves. That does have an impact on what the usual care group then goes and gets after they enter the study. Even though the usual care is unprotocolized, it certainly is restricted by the information that is given to the individual who would have to accept randomization to any of the arms. In your discussion of the wild type versus protocol-driven usual care, that first step of a patient agreeing to participate in the trial does affect usual care. Maybe that is what you meant earlier about a risk associated with usual care because you are in the study and that influences in some way what you seek or get. That issue is important for participating in the clinical trial.

Dr. Weinstein: The question of usual care in surgery is different. To a surgeon, it might be that my operation is usual care for a given problem and now a new thing comes along. Whether operative or non-operative, do I want to compare this new "thing" to a nonsurgical treatment or to the latest surgical treatment? I am not really interested in the nonsurgical usual care. Most surgeons feel more comfortable comparing surgery A with surgery B because there is no other option. Then what about pain injections – this would require a different design of the trial to answer that question about that specific modality. In surgery, the definition of usual care would be surgery. To a surgeon, that means doing Operation A versus Operation B. When a new technology comes along, they do not necessarily want to compare that to non-operative "watchful waiting," which may be better but it will never be tested.

Dr. Thompson: You are still looking for some help on this protocolized notion. An example is that we first make the decision about the bounds of usual care. The scientific question requires some comparison to usual care, so we identify the bounds. Then we look within usual care for explained and unexplained variation. If we can explain the variation (and it is not necessarily important that it be evidence based but it must be explainable), then our approach, in a consensus process, has been to protocolize that. For example, we do not know how many days to treat patients with ventilator-associated pneumonia with antibiotics so there is a lot of practice variation. One approach would be to agree that it is unreasonable to give them 1 day and it is unreasonable to give them 1 month, so then we look at how clinicians customize within that bracketed range of competent care. If a broad-enough group of clinicians representing this broad standard can agree on a set of rules on which to customize, you can protocolize usual care. The advantage of that approach is that you know what it is and you can make some inferences about something that is different than usual care. The problem is that it is really an ideal and it is difficult to get experts to agree. But it is a worthy goal and a necessary exercise when usual care comparisons are relevant to the scientific question.

Dr. Silverman: I do not like the term “wild type” – it sounds pejorative.

Regarding protocolized care, there may an interest in ensuring that all the doctors are using the dose shown to be effective in previous trials, rather than having physicians use a substandard dose and then you would have an inferior usual care group. It could be simple in terms of mandating the dose and duration, and you could probably get physicians to agree on that. Then there is protocolized care in which you want to affect design issues and where you want to protocolize a common practice among a particular group of physicians. That might or might not be difficult to do. Some critical care studies have been able to protocolize the control group.

Dr. Friedman: Most of the time when we randomize people to one group or another, we are talking about randomizing people to strategies. It is not Treatment A versus Treatment B, where B may be usual care or not, but it is a strategy. Not everyone will take the drug or will get the surgery or will get “wild type” usual care. In some sense that is protocolized because we start with this approach and, if necessary, we will follow it up with another approach. We treat people as we need to treat them if what is being done to them is good, bad, or indifferent. The same applies to the usual care group. It is a strategy because we can protocolize it. We say, “You start this way and if that works, great. If it does not work, there is a second or a third approach.” That is commonly done to allow for a variety of approaches to treatment.

Continued Discussion of Ethical and Scientific Considerations

Dr. Fleischman: Our goal during the next 1.5 hours will be to move this discussion toward thinking about stakeholders and those to whom we might want to give points to consider about their responsibilities or their actions – including investigators, scientific reviewers, IRBs, and potential participants.

We have added to our panel Alex London to enhance our philosophical weight. We wanted to recap on one issue that came up during the break. How do we justify scientifically the broad variation in usual care and the potential that usual care becomes individualized or customized care because it is out in the community and individual physicians or clinicians can dramatically impact on that and individualize it? From a scientific perspective, how do we justify the broad variation, customization, and patient preference? From a statistical or design perspective, are these surmountable? Are immense numbers needed because of the broad range and complexity of the variation that would occur in these trials?

Dr. Friedman: Numbers are good, but they do not solve the problem of a lousy design. You can have all the numbers you want, but unless the comparisons are valid, among the various possible approaches, it will not help interpret the data. The issue of “take every possible approach in usual care” comes back to some things discussed earlier – what kind of evidence do you have that they are poolable and describable in some meaningful way? “I have shown that my new treatment is better than what?” How does that help the clinician?

Dr. Dixon: You can tolerate quite a range. Here is a made-up example: if usual care is effectively palliation, there are all kinds of ways to meet that objective. At the end, if the trial is evaluating something else compared to that, it would be easy to say at the end that the new approach is either better than palliation or not. In this kind of case, it is not a problem to have a wide range of options within the usual care arm, provided there is the restriction that rules out those options that are already known to be inferior.

Dr. Fleischman: How do we rule them out? In the design piece, as an investigator, how do we determine what is in and what is out? Dr. Silverman did not like the term "wild-type" usual care, but I did not mean that in his colleagues' ICUs. What I meant is, in the Great Beyond of all doctors doing all things to all people. How do you rule them in or out?

Dr. Dixon: In some cases there would be preexisting evidence from clinical research. People learn things by analogy or by making appeals to the same drug class. There are many ways people can learn without doing a full-scale clinical trial. It is possible to identify approaches that are clearly inferior.

Dr. Fleischman: Do we have to do that?

Dr. Levine: You do not. If you decide to use usual care as a comparator, you are not trying to make a statement that your new or old intervention is better than each and every component of usual care. It is whether your new surgical or pharmaceutical intervention has a better outcome than people can find on their own when seeking usual care in their own communities. How does this differ from historical controls? At the outset you are putting people through certain inclusion and exclusion criteria so you know some of the baseline attributes of those who are either getting randomized to surgery or to seek usual care. We have to acknowledge that usual care in the community is not everything we would like it to be. But there are certain sorts of monitoring functions that limit the possibility of introducing truly incompetent therapies into usual care. Of interest to us are the FDA and malpractice litigation that will enforce the requirement not to go below the standard of care.

Dr. London: We have to be careful here. If we are defining usual care in terms of what type of intervention someone might receive from a local clinician, it is important to ask how many of those practices are retrograde practices that would have been transformed if the subpopulation of practitioners (which could be very small) had been informed by recent research. You want to distinguish elements of usual care that are vestiges of practices that should have died out from elements of usual care where you have a reasonable minority of clinicians with some reasonable hypotheses about why their interventions might work. Those are only two examples in which you may not have the same attitude toward what is "out there" in usual care.

Dr. Fleischman: But as an investigator, how do I do this? If I have a usual care arm that is the natural history of how this problem is being managed in the community, what are my obligations to find out how it is done, putting boundaries around those, or is this just a matter of comparing my new intervention to what people are getting in the

community? Do I have to worry about that or am I just a natural observer of the natural history, in that arm of the trial, as compared to my experimental treatment?

Dr. Silverman: If the evidence is compelling enough from a human subjects safety point of view and a design point of view and if the statistical thought is compelling enough, you do need to put a boundary on usual care and you would operationalize that in your exclusion criteria.

Audience: The Canadian Clinical Trials Group frequently begins their assessments with a survey to identify the attributes of usual care in order to conduct the design in a manner that is scientifically valid. We should point out that, frequently, what people say they do and what they actually do are different.

A comment for the panel: there is the observation that only a small number of clinicians actually exhibit compliance with evidence-supported and expert opinion derived guidelines and recommendations from acknowledged societies and prestigious organizations. Is the observation that, in clinical practice, compliance with recommended practice is very low *prima facie* evidence of incompetent care? This issue was raised earlier – that incompetent care has to be eliminated.

Dr. Fleischman: So if there is an evidence base and learned groups believe that evidence base is powerful but practitioners are not using it, what are our obligations in the study that encompasses all that substandard practice versus the guidance-level practice as compared to our new intervention? How do we deal with that? Do we have to put boundaries around that?

Dr. Friedman: I think so. If there is some intervention out there that is thought to be useful but, for whatever reason in a number of cases, it is not being used and you do a study to compare a new intervention against what is out there (which is a mixture of 25 percent of the people using the presumed beneficial usual care and 75 percent not), what does it mean to come up with an answer that your new intervention is better than usual care? You cannot inform medical practice much in that setting. You do want to put some bounds on what the usual care is.

Dr. Fleischman: We are not talking about usual care when we do this. We are really talking about a structured evidence-based care against a new care. Even the language of usual care may not work here.

Audience: The proposition is that, when trying to determine what components of care might constitute the usual care group, you need to make that decision based on the best appraisal of evidence available, which can come all the way from basic science through antecedent trials, animal trials, and human trials. The people best placed to make those kinds of judgments have to already be familiar with the intellectual field of interest. You must rely on experts to appraise the evidence in a systematic fashion, and it should be a public appraisal. When they come to a determination that these components represent a usual care group, one can make a judgment about the process and the determination they have reached.

Dr. Fleischman: That is one view about how usual care should be determined, but it is not the only approach. We see a large number of research designs in which usual care is much more broad and is only one measure of what is in the "wild." We are talking about asking better questions and putting better boundaries on usual care and thinking about the evidence.

Going back to Dr. Weijer's original comment in which he laid out the methodology, design is primary. He was arguing for a sequential approach to design.

Dr. Weijer: I made three points. The first point was that, in all cases, the scientific question ought to drive the trial design. Following from that, a clear and justified question drives the need for a usual care comparator in a particular case. Also following from that is that the usual care arm answers the clear and justified scientific question in a way that a protocolized care arm cannot.

Dr. Silverman: The presence of diversity does not necessarily imply incompetent care. The absence of evidence does not necessarily imply evidence of bad care. If there is no evidence – and frequently all we have is expert opinions – that is good usual care.

Dr. London: It is important to distinguish the kind of diversity you might find in usual care. If you decide that you need a usual care arm in your clinical trial, the next question is whether the boundaries of usual care in the community fall within the boundaries of competent care or care that a researcher would feel OK about study participants having access to.

For the set that remains, what do the interventions look like? Is there a plurality of discrete intervention or is there a "smoosh"? I do not think we want to confuse several discrete sets with a "smoosh," because they give you very different questions related to whether you want to choose one or two of the discrete sets as the comparator.

Dr. Fleischman: Are there other comments about investigator obligations in considering usual care trials?

Dr. Thompson: To echo the comment about access to care in the usual care group, some of these trials that use unrestricted usual care are really health policy trials. Kaiser might say in one of their health plans they will introduce a package of asthma therapies, which will include inhaled steroids, education, peak flow meter, etc. The question is not which one in the bundle is improving outcomes but that that whole approach might be better than usual care. If you are testing a package of asthma education in an HMO versus asthma education in an urban asthmatic population, in a setting where you have a prior suspicion that 90 percent of the patients do not have any access to health care, that is not an acceptable usual care comparison arm for those policy-like trials. There is an obligation to explore what usual care is. It is a scientific question and you certainly would not randomize to a group that you *a priori* did not understand.

Audience: Does the investigator have the responsibility of defining the scope of usual care? What if only the "blue states" are doing a particular protocol? Do I only have to see what is usual care for the "blue states"? If it is only Massachusetts, is that usual care? If it is only using the ICUs at one hospital, then I have 15 different ICUs and none of them talk to each other and all have different usual care. There is responsibility there. It would be different if it were an investigator sponsored single-site trial versus a multisite trial.

Dr. Thompson: This is an important question. If you design a national clinical trial and you agree this is a nationally important question – in a national sense we do not know if A or B is best – then is there an obligation to make sure that when you are judging incremental risk over not participating in the trial (i.e., usual care), is it some local IRB responsibility to judge incremental risk over A or B? It might mean, in an extreme example, that if usual care in Texas is way outside, the incremental risk equation at the local level might be such that that trial is uncomfortable. What is the level of evidence that the IRB has to have to make the incremental risk assessment of A over B at that local institution? Mechanisms are in place at some institutions to do this better than others. One of our institutions has a group of partner institutions and gives the IRB advice about incremental risk. When we submit a trial to our local IRB, even if it is a national trial, we are required to inform our IRB about incremental risk. Fundamentally that is a comparison about usual care at the institution and participation in the trial. The level of evidence required to make that judgment may not need to be prescriptive; most of the time it will be quite qualitative.

Dr. Silverman: The answer is dependent on the question. If the question is whether your intervention is better than usual care, then you do have an obligation to be representative of the U.S. population. For example, the SPORT trial focuses on community and academic institutions.

Dr. Thompson: What if it is A versus usual care? Now the study is fundamentally addressing different questions in different institutions. That is a whole different level.

Dr. Fleischman: There are two issues. One has to do with the national representativeness of a sample – that is a scientific question that should be raised, and most studies do not represent a national sample but usually represent a series of local samples. Sometimes you can generalize and sometimes you cannot. Within that individual institution there are potential incremental risks based on these issues – those are different scientific questions on which the IRB needs to have sufficient information or capacity to generate that information. It is always the investigator's responsibility to educate the IRB. The problem is whether the IRB is educable and whether there are enough outside experts – whether the IRB has the wherewithal to know what it does not know.

Dr. Swanson: I was thinking about your question about the investigator's obligation and how that relates to the usual care group—what is necessary when people are randomized to multiple groups?—and that our obligation for the modalities we are using need to be based on empirical information. There are lots of other interventions for

which we do not have that information. It seems that our obligation is to inform people about the empirical basis for interventions associated with a particular condition. Usual care may or may not have that empirical basis. It is the obligation of the investigator who is conducting the trial to give everybody that information – what has an empirical basis and what does not. When there is randomization to a usual care group, participants in the trial would be given information about what treatments the investigators consider to have an empirical basis for treating that medical condition. Participants are free to choose the other one without the empirical basis, for whatever reason. I do not quite see how the investigator obligation goes beyond the information that must be provided for all subjects who must decide whether to accept randomization to usual care or to one of the other treatments.

Dr. Fleischman: There are some things in the field of developmental and behavioral psychology, particularly for vulnerable children with disabilities, that some people think are risky or hurtful or expensive, that people get all the time in various parts of the country. Are you suggesting that you remain neutral about things that you, as a professional in your field, believe have no evidence base and are merely "charlatanism"?

Dr. Swanson: The attitude might be, "I want to do things until you prove it is wrong." But that is not the way we do things.

Dr. Fleischman: You are the investigator, coming to me as the IRB chair, and you are going to randomize subjects. Some will go into different arms of your clinical trial and some will go out into the community. You are going to be following some into the community to find out what happens to them. What obligations do you have to share with them the variability of what is out in the community?

Dr. Swanson: If there is evidence of harm, then that needs to be shared. If there is an empirical basis for intervention, that also needs to be shared. A critical issue is, what about those interventions for which there is no critical evidence, pro or con? What do you do in that case? I feel I should not say, "There is no evidence, so do not do that." It is arrogant of me to do that. I feel much better when I am not certain and there is some empirical information, by whatever criteria, and I say, "Here is what people have shown to be effective in certain circumstances; go consider that." In the middle ground, where there is no evidence, many people will say, "Prove me wrong and I will stop doing it, but I will continue until then." I do not know what to do about that.

Dr. Fleischman: My concern is, as we (at the IRB) look at this protocol and as you (the investigator) recruit subjects, if you are going to measure these different kinds of treatments out in the community, are we (at the institution reviewing your study) promoting a misconception among the subjects that we believe that usual care, as delivered, is an acceptable way to go? Are we sending a message to the community for which we are responsible, since we are reviewing your proposal? It is part of my concern about the wild-type/usual care strategy that we have some obligations around boundaries, or at least truth telling, that goes beyond the level of evidence when we are

talking about Laetrile or interventions for disabled children that are mostly "wallet biopsies."

Dr. London: I want to distinguish two cases; I am not sure whether we are conflating or not. When the usual care arm is what you might call "community care" – when you are following them into the community where they get whatever is available – the limits of your reach are informing subjects but you cannot prevent them from going to a charlatan. I want to contrast that with the case in which the usual care is administered by other researchers who have agreed to be included in the trial. There you do have an obligation to make sure that the other researchers participating with you in the trial represent what you think of as reasonable models of care or reasonable practice. When you choose the people with whom you collaborate, you have more control over that than in the case where people can seek whatever care they want.

Dr. Friedman: Yes, but you still have to have a study that addresses a sufficiently important question. If you have some notion that a fair number of the people in the usual care arm will get treated by a charlatan, the IRB has an obligation to question the validity of your study's conclusions. You do have at least some obligation to make an estimate of what proportion of the people will go in that direction and how you will handle it.

Dr. Pater: There was a time when there was great controversy about whether any form of cytoxic chemotherapy was beneficial in patients with advanced lung cancer. Eventually those studies were done. They were randomized trials of a form of chemotherapy versus what we called "best supported therapy." It was basically anything but chemotherapy. If those trials were done today, most of the patients in the best-supported-care arm would have been taking some other complementary or alternative therapy. We would not have told the experimental arm subjects that they could not take complementary therapies. So we would be saying, chemotherapy plus whatever else versus whatever else, and we would have a fair comparison of chemotherapy to no chemotherapy. In at least some of the designs we are talking about, you are not excluding access to the other forms of usual care in the experimental arm, yet in other designs you are. One needs to make that distinction. In the nonsurgical arm of the SPORT trial, people could go out and get all those other things, but they could not have surgery. That is a clear question; it is surgery versus no surgery in the background of getting usual care. But if the trial design consists of one arm where you cannot have usual care versus another arm allowing usual care, that is a different question.

Audience: The question is whether usual care should be an arm in a research project. What seems to be the question now is, how do you determine what usual care is? Having that arm, in some cases such as surgery, it might be difficult to determine what the usual care arm is other than surgery. For example, testing a pharmaceutical drug for high cholesterol, statins are a proven therapy. How do you move science and medicine forward if you do not compare new products to what is considered usual care or what is scientifically proven effective for that ailment?

Dr. Fleischman: No one is arguing that you should do that.

Audience: I am getting concerned and confused because I am hearing discussion of randomization to an arm that is usual care, yet we had at least two studies that showed that if you did not want to participate in the randomized trial, you could agree to be in an observational study that involved tracking usual care. There would be a different level of evidence and different considerations on the part of the IRB in those two situations.

Dr. Levine: Why does the IRB feel a sense of responsibility for what people get in usual care? Before someone comes to the IRB and asks to evaluate a new intervention, everyone in the IRB's domain is getting usual care, including some care administered by charlatans. The reason that the IRB usually considers that it has a responsibility for the control group is that the enrollment in the control group will be to follow a plan or regimen or accepted therapy that is recommended by the investigator who is part of the IRB's institution. Of course they have some responsibility to determine what that is and whether or not it really can do what it is alleged to do.

In the case of a usual care comparator, this is not something recommended. This is saying, "You are eligible to be in our clinical trial but according to our randomization scheme you are not going to get the intervention; you are going to go back out and do what you would have done otherwise." That is a different set of circumstances.

I want to raise the issue of what you do if you know there is someone in the community who is offering charlatan services – doing something we all know does not work. I think that is a minority situation and I do not think we should alter our thinking about the *usual* usual care arm by focusing on this minority, but just in case. For example, what if I am practicing in a small town in Tennessee where some guy is making personalized antibodies to cancers. I might make as part of the protocol requirement at this site in Tennessee that "usual care means usual care minus personalized antibodies." That would be a slight move away from usual care, but you could specify that in your description of the study.

Dr. Weijer: I will try to finesse a number of the answers that have been given. Clearly, the IRB has some challenging questions to consider when reviewing a study that involves a usual care control. The first and most important question is not one of consent, it is whether we have a valid study – whether a clear and justified question is asked and whether the trial is designed so that question will be answered reliably. Secondly, the IRB needs to ask itself whether the question is valuable and whether the value of the study is such that it counterbalances the nontherapeutic risks to subjects, if such risks exist.

Dr. Levine is right – IRBs have no jurisdiction over the physician-patient relationship. Dr. London is right – IRBs do have oversight over the researcher-subject relationship. So it will depend whether the usual care arm in whatever form actually invites in community practitioners as researchers in the study. That would be a more controlled usual care arm than a take-all-comers arm. In that more controlled circumstance, the

IRB would want to make sure that all treatments being delivered are consistent with competent medical care.

The question that bothers me most is how often it will be that good science demands a control arm that is unrestricted, take-all-comers, usual care. That is what I am having the most trouble with. I am relatively skeptical of the utility of that arm, for reasons that a number of panelists have indicated.

Audience: Research misconception is one of the risks the IRB has to look at with a usual care control arm. You are not just thanking participants for being in the usual care control arm, you are following them (otherwise the arm would be useless). You now have information about them. While you may have said, "So what if they are doing prayer for appendicitis?" But now you know about it. As a researcher, you now have more information and more responsibility. That is a risk.

Dr. Fleischman: Where I would disagree with Dr. Weijer's characterization of Dr. Levine's view is that the IRB has some concerns when the participant may inappropriately think we have some umbrella that says we accept all of that physician-patient relationship interaction, and all is acceptable because we are observing it. It is a big leap to believe that our participants understand the difference.

If Dr. Swanson is in the business of studying ADHD and figuring out the best ways to treat that problem, and if participants in Dr. Swanson's overall research are not getting any of his fine alternatives that are well protocolized, then my concern as an IRB member is that participants who are not getting the protocolized intervention need to understand that. Participants in the observation group need to understand that they are in the observation group and are not receiving the treatment Dr. Swanson is testing, related to the study's title. That is a challenge in the informed consent world that we do not meet well.

Dr. Weijer: How does that undermine my point? Fundamentally what I was talking about was the IRB's purview vis-à-vis physician-patient relationships and researcher-subject relationships, and I was suggesting that both Dr. Levine and Dr. London characterized those correctly. If there are systematic misunderstandings of research subjects above and beyond harm-benefit determination by the IRB, then they need to be disabused of those systematic understandings. That is a challenge for the consent process, not a challenge for the harm-benefit determination. Similarly, it is not a call (at least in Dr. Levine's case) to restrict the usual care arm.

Dr. Levine: I would disagree with these last two points. First, it is not as if I am seeing a use for a usual care arm in the majority of clinical trials. I think that the need for a usual care arm would apply to a minority of clinical trials. When it does apply, one of the first rules is that employment of the usual care arm must assist in gathering evidence considered essential.

On the point of whether or not the IRB should feel a responsibility for what goes on in usual care, even to the extent that you suggested, I will press this to an extreme.

Imagine the IRB reviewing the research proposal submitted by an ethnographer – where it is their business to find out what is going on in the world. Many ethnographers and anthropologists have written about the extreme tension they experience of witnessing something that they believe is wrong. Many papers have been written from anthropologists who have worked in cultures where they see things that make them say, "This is evil but it is not my role here to do something; it is my role to come away from this situation and describe the culture." When comparing something to usual care, we cannot go that far. So if we have some "crackpot" making personalized antibodies to cancer, we can disinclude that. Some people may feel the same way about bariatric surgery. But then what remains is no longer strictly a usual care comparison.

Dr. Fleischman: We are talking about a therapeutically intended intervention and a usual care observational arm. I am more concerned about those trials than an ethnographer or a purely observational study, because participants are being recruited into this trial that has a therapeutic intent or a study of different kinds of interventions of clinical relevance. That is my overarching concern.

Dr. Levine: So what you can do is build into the process and forms documenting informed consent that say, "We are not to be held accountable for the practice of medicine outside of this institution."

Dr. Fleischman: What I would like to do with our remaining time is to ask the following question: If you were given 3 minutes to tell either the sponsors of research at the level of the NIH or the regulators of research at the level of the FDA and the OHRP about usual care in clinical trials, what would that take-home message be?

Dr. Friedman: Let me separate the kinds of usual care. I am not talking about usual care as background therapy for everyone; I am talking about usual care as a separate arm compared with some protocolized care. In that sense, all I can do is repeat what a number of people on the panel have said: What is the question, and is that question answered best by having a broad spectrum of care as a comparison or by having a more protocolized, narrowly focused strategy? It comes down to the justification for the question you seek to answer.

Dr. Levine: I agree with Dr. Friedman.

Dr. Pater: I would make sure that they go through a definitional exercise, so people are talking about the same thing when they discuss usual care. We would have to be very careful in adopting a design that was known not to have an evidence-based standard "out there." I do not see how this form of research design can apply when there is an evidence-based standard that could be used. In most circumstances, it should be possible to identify one or two forms of relatively well-specified care that most people would agree that nothing is better than. Therefore, when the trial is over and if the new treatment is better than that, we are happy with the results.

Dr. Silverman: In addition to Dr. Friedman's comments, one of the take-home messages is that usual care could be acceptable to do. Sometimes the language we

use imprisons us, and that was the basis for my not wanting to use terms like “wild type” or “haphazard care.” In other discourse, it might prejudge the endeavor to use terms like “wild type.” One can learn a lot from having a usual care group and one could miss opportunities by not having a usual care group, depending on the question being asked.

Dr. Swanson: The information on interventions that might be out there in usual care, that have some empirical basis either in favor or against them, should be provided to all participants who are subject to randomization to the treatment or the usual care groups. That seems to be the essential feature. I do not understand the value of the usual care group when randomization is not part of the design.

Dr. Swart: I would encourage supporters of research and the regulators to make sure protocols specify the research questions and make sure that usual care fits the purpose. Secondly, where possible, define minimum standards of usual care. Thirdly, encourage investigators to be more explicit in terms of specifying all the things to do with usual care – for example, selecting investigators’ restrictions on treatments that may not be allowed and standards of followup.

Dr. Thompson: I echo what has been said before. This is fundamentally a scientific question, not an ethics question; the ethics follow the science. From an ICU standpoint, there are many unanswered questions. For ICU trials, most but not all trials should and can answer important questions without an unrestricted usual care arm. As the field of ICU medicine matures, most may change. Most trials should not be using unrestricted usual care to answer important questions. Maybe as we have an evidence base and usual care becomes more defined, that will become a narrower and more reliable comparator. This is dynamic and relates to the state of knowledge in 2005.

Secondly, advice to regulators. When you are comparing A versus B to usual care in your institution and trying to make the incremental risk assessment, regulators should allow judgment in this domain and should not require strict mathematical descriptions of usual care and an overly deterministic methods for making this comparison. We have to be very careful here and allow for some judgment from expert clinicians in the community about where an A and B fall in relation to usual care in that incremental risk assessment.

Dr. Weijer: As others have said, the scientific question needs to drive the trial design. The use of a take-all-comers usual care arm needs to be driven by a clear and justified scientific question. The only type of question that could justify a take-all-comers usual care comparison arm would be a pragmatic question. Therefore, it is interesting that this question has come to the fore as the result of an explanatory trial where it was suggested (wrongly) that a usual care arm would either improve the science or the ethics of that trial, or both.

I am going to put this statement out to be refuted: No unethical two-arm comparison can be made ethical by the addition of any third arm. If the ethics of the trial were the issue, then seeking to add a third arm is ill conceived from the start.

Dr. London: I am sure I would want to object to that but I do not know why! I agree with just about everything that has been said, prior to Dr. Weijer, and also most of what Dr. Weijer said.

If I had a chance to talk to those stakeholders who are not IRBs, I would want to distinguish between these various stakeholders. IRBs function often as a filter. A lot of questions are asked and answered before they get to the IRB and the IRB sees these things later on. I would want to point out that, in this forum, a lot of issues have been raised that deal with evidence-based medicine, issues in the diversity of care in practice that are salient in the discussion about usual care arms in clinical trials. I would want to ask those stakeholders what we are doing to ensure that the high-quality research we conduct is translating in the medical community to changes in practice. That is the big "pink elephant" in the middle of the room. In some cases, more emphasis on that will also reduce the need for this kind of trial design.

I want to make sure we do not overlook the case in which there is an important need to transfer from the point where we have a "smoosh" to the point where we have more discretely defined competing interventions or treatments that we can then subject to different kinds of trial designs. SPORT is a good example of a way in which this kind of usual care group might provide a valuable role in moving from the "smoosh" to more discrete comparators.

Dr. Dixon: I would like to reemphasize a few points. The role of usual care arm clinical trials is exactly in a situation in which no consensus has been formed in the medical community as to what the right approach or approaches might be for a particular situation. The very fact of doing research with usual care will change some of the features of usual care – for example, scheduling followup visits. It is the Hawthorne Effect – some disturbing of usual care will occur in any circumstances. The fact that a clinical trial is conducted with a usual care arm may accidentally change the specific meaning of that usual care. For instance, there may be a range of alternatives but somehow in the particular clinical trial not all of those alternatives show up at once. What is in the usual care arm at the end differs in some way from what was anticipated in the study's design, and that is bound to have consequences for interpretation of the results.

Final Thoughts

Dr. Fleischman: I want to thank the Office of the NIH Director for putting together this panel and all the speakers, and for addressing a serious and important question in the research community. Thanks particularly to Liza Dawson and Allan Shipp, who put together much of the work that went behind this elegant activity.

Dr. Patterson: I want to thank the audience for your fortitude and participation. Your input will provide influence and wisdom into the work products that will come out. I also want to thank our esteemed speakers and panelists. This has been a tremendous

dialogue on a very important question. Your candor, your diversity of opinion, and the respect that you showed for each other and the audience has been greatly appreciated.

There will be a meeting proceedings. A record of the meeting will be produced. A CD of the meeting will also be produced. We would like to push the envelope a little farther. As you know, we have the points to consider document in the notebooks and we are going to be working on refining that, incorporating some of the concepts that have emerged and the moments of illumination that have occurred here, which we hope will be reflected in that points-to-consider document. This is not likely the last event of its sort.

As many answers as we came up with, we have raised an equal or greater number of questions. For instance, we will be looking at the issue of standard of care/usual care in trial design on the international scope. Look for a CD, a meeting proceedings, and ultimately an article of the conceptual framework presented here.

Conference adjourned.

Appendix

Considering Usual Medical Care in Clinical Trial Design: Scientific and Ethical Issues

Bethesda Marriott Hotel, Bethesda, Maryland ♦ November 14-15, 2005

AGENDA

November 14, 2005

8:00 a.m. - 8:30 a.m.	**Registration**
8:30 a.m. - 8:40 a.m.	**Welcoming Remarks** **Raynard S. Kington, M.D., Ph.D.** Deputy Director National Institutes of Health **Amy P. Patterson, M.D.** Director Clinical Research Policy Analysis and Coordination (CRpac) Program Office of Science Policy Office of the Director National Institutes of Health
8:40 a.m. - 9:00 a.m.	**Introduction to the Conference** **Robert J. Levine, M.D.** Co-Director Yale University Interdisciplinary Center for Bioethics Professor of Medicine and Lecturer in Pharmacology Yale University
9:00 a.m. - 9:40 a.m.	**Variation in Medical Practice and Its Implications for the Design of Control Arms in Clinical Trials** **John E. Wennberg, M.D., M.P.H.** Professor Departments of Medicine and of Community and Family Medicine Dartmouth Medical School
9:40 a.m. - 10:20 a.m.	**Federal Agency Views on the Relevance of Usual Care Control Groups (FDA, NIH, AHRQ, OHRP, CMS, VA, CDC**) (5-minute presentation for each, with 15-minute Q&A)
10:20 a.m. - 10:40 a.m.	**BREAK**
10:40 a.m. - 11:20 a.m.	**Design Considerations in Randomized Controlled Trials With a Focus on Usual Care Arms** **I. Basic Principles of Clinical Trial Design** **Janet Wittes, Ph.D.** President Statistics Collaborative, Inc.

11:20 a.m. - 11:50 a.m. **Design Considerations in Randomized Controlled Trials (continued)**

II. FDA Perspectives on Usual Care Control Groups in Regulatory Decisionmaking

Robert J. Meyer, M.D.
Director
Office of Drug Evaluation II
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

11:50 a.m. - 12:40 p.m. **Ethical Considerations in Randomized Controlled Trials: Focus on Usual Medical Care and Ethics of Trial Design**

Charles Weijer, M.D., Ph.D.
Associate Professor
Department of Philosophy
Talbot College
University of Western Ontario

12:40 p.m. - 1:40 p.m. **Luncheon Speaker—Challenges of Practicing Evidence-Based Medicine: Integrating Science and Clinician Experience in Patient Care**

R. Brian Haynes, M.D., Ph.D.
Professor and Chair
Department of Clinical Epidemiology and Biostatistics
Faculty of Health Sciences
McMaster University

1:40 p.m. - 2:20 p.m. **The Acute Respiratory Distress Syndrome Network (ARDSnet): Lessons Learned for the Design of Critical Care Research**

B. Taylor Thompson, M.D.
Director
Medical Intensive Care Unit
Medical Director
ARDS Network Clinical Coordinating Center
Massachusetts General Hospital

2:20 p.m. - 2:45 p.m. **Discussion of ARDSnet**

2:45 p.m. - 3:00 p.m. **BREAK**

3:00 p.m. - 6:00 p.m. **Case Studies**

3:00 p.m. - 3:30 p.m. **Case Presentation: Case Study #1**

International Collaborative Ovarian Neoplasm (ICON) Trials of Ovarian Cancer Treatment

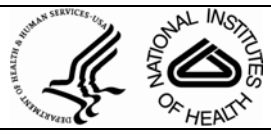
Ann Marie Swart, M.R.C.P., M.Sc.
Senior Clinical Epidemiologist
Cancer Division
Clinical Trials Unit
UK Medical Research Council

3:30 p.m. - 3:45 p.m.	**Commentary on Case Study #1**
	Joseph L. Pater, M.D., M.Sc. Edith Eisenhauer Chair in Clinical Cancer Research Director National Cancer Institute of Canada Clinical Trials Group
3:45 p.m. - 4:35 p.m.	**Panel Discussion of Case Study #1**
4:35 p.m. - 5:05 p.m.	**Case Presentation: Case Study #2**
	Multimodal Treatment Study of ADHD (MTA)
	James M. Swanson, Ph.D. Professor Department of Pediatrics Director Child Development Center School of Medicine University of California, Irvine
5:05 p.m. - 5:20 p.m.	**Commentary on Case Study #2**
	Julie Magno Zito, Ph.D. Associate Professor of Pharmacy & Psychiatry Department of Pharmaceutical Health Services Research University of Maryland, Baltimore
5:20 p.m. - 6:00 p.m.	**Panel Discussion of Case Study #2**

November 15, 2005

8:00 a.m. - 8:30 a.m.	**Registration**
8:30 a.m. - 9:00 a.m.	**Case Presentation: Case Study #3**
	Spine Patient Outcomes Research Trial (SPORT)
	James N. Weinstein, D.O., M.S. Professor and Chairman Department of Orthopaedic Surgery Dartmouth Medical School
9:00 a.m. - 9:20 a.m.	**Commentary on Case Study #3**
	Steven N. Goodman, M.D., Ph.D. Associate Professor of Oncology, Pediatrics, Epidemiology and Biostatistics Division of Biostatistics Department of Oncology Johns Hopkins Medicine and Johns Hopkins Bloomberg School of Public Health
9:20 a.m. - 10:10 a.m.	**Panel Discussion of Case Study #3**
10:10 a.m. - 10:30 a.m.	**BREAK**

10:30 a.m. - 12:00 noon	**Roundtable Discussion: Development of Ethical and Scientific Principles To Guide Considerations of Usual Medical Care in Clinical Trial Design** **Roundtable Chair: Alan R. Fleischman, M.D.** Chair Federal Advisory Committee Ethics Advisor National Children's Study National Institute of Child Health and Human Development National Institutes of Health
12:00 noon - 1:00 p.m.	**LUNCH**
1:00 p.m. - 2:30 p.m.	**Continued Discussion of Points to Consider**
2:30 p.m. - 2:50 p.m.	**BREAK**
2:50 p.m. - 4:00 p.m.	**Final Discussion and Summary**
4:00 p.m.	**ADJOURNMENT**



NIH Program on Clinical Research Policy Analysis and Coordination

Considering Usual Medical Care in Clinical Trial Design: Scientific and Ethical Issues

PLANNING COMMITTEE*

**Affiliations, position titles, and addresses of Planning Committee members reflect their positions as of 2003,when the Committee was assembled.*

Duane F. Alexander, M.D.
Director
National Institute of Child Health and Human Development
National Institutes of Health

Jonathan (Josh) Berman, M.D., Ph.D.
Director
Office of Clinical and Regulatory Affairs
National Center for Complementary and Alternative Medicine
National Institutes of Health

Carolyn M. Clancy, **M.D.**
Director
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services

Ezekiel J. Emanuel, M.D., Ph.D.
Chair
Department of Clinical Bioethics
Warren Grant Magnuson Clinical Center
National Institutes of Health

Ellen G. Feigal, M.D.
Acting Director
Division of Cancer Treatment and Diagnosis
National Cancer Institute
National Institutes of Health

Lawrence M. Friedman, M.D.
Consultant
Former Director
Division of Epidemiology and Clinical Applications
National Heart, Lung, and Blood Institute
National Institutes of Health

John I. Gallin, M.D.
Director
Warren Grant Magnuson Clinical Center
National Institutes of Health

Saul Malozowski, M.D., Ph.D., M.B.A.
Senior Advisor for Clinical Trials and Diabetes Translation
Division of Diabetes, Endocrinology, and Metabolic Diseases
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health

Peter J. Mannon, M.D., M.P.H.
Chair
Institutional Review Board
Laboratory of Clinical Investigation
National Institute of Allergy and Infectious Diseases
National Institutes of Health

Joan A. McGowan, Ph.D.
Chief
Musculoskeletal Diseases Branch
Extramural Program
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institutes of Health

Amy P. Patterson, M.D.
Director
Clinical Research Policy Analysis
and Coordination (CRpac) Program
Office of Science Policy
Office of the Director
National Institutes of Health

Marcel E. Salive, M.D., M.P.H.
Director
Division of Medical and Surgical Services
Coverage and Analysis Group
Office of Clinical Standards and Quality
Office of the Administrator
Centers for Medicare & Medicaid Services

Bernard A. Schwetz, D.V.M., Ph.D.
Acting Director
Office for Human Research Protections
U.S. Department of Health and Human Services

Belinda Seto, Ph.D.
Acting Director
Office of Extramural Research
Office of the Director
National Institutes of Health

Lana R. Skirboll, Ph.D.
Director
Office of Science Policy
Office of the Director
National Institutes of Health

David Shore, M.D.
Associate Director for Clinical Research
National Institute of Mental Health
National Institutes of Health

Robert J. Temple, M.D.
Associate Director for Medical Policy
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

Deborah A. Zarin, **M.D.**
Director, Technology Assessment Program
Agency for Healthcare Research and Quality

DRAFTING COMMITTEE

Points to Consider Document

Ezekiel J. Emanuel, M.D., Ph.D.
Chair
Department of Clinical Bioethics
Warren Grant Magnuson Clinical Center
National Institutes of Health

Lawrence M. Friedman, M.D.
Consultant
Former Director
Division of Epidemiology and Clinical Applications
National Heart, Lung, and Blood Institute
National Institutes of Health

Steven N. Goodman, M.D., Ph.D.
Associate Professor of Oncology, Pediatrics,
Epidemiology and Biostatistics
Division of Biostatistics
Department of Oncology
Johns Hopkins Medicine and Johns Hopkins Bloomberg School of Public Health

Deborah A. Zarin, **M.D.**
Director
Clinical Trials.gov
Assistant Director for Clinical Research Projects
National Library of Medicine
National Institutes of Health

Liza Dawson, Ph.D.
Health Science Policy Analyst
Clinical Research Policy Analysis and Coordination (CRpac) Program
Office of Science Policy
National Institutes of Health

Bimal Chaudhari
NIH Summer Internship Program Student
NIH Office of Biotechnology Activities
MD/MPH Candidate - Boston University
MED/SPH '09

Participating CRpac Staff Members

Amy P. Patterson, M.D.
Director
Clinical Research Policy Analysis
and Coordination (CRpac) Program
Office of Science Policy
National Institutes of Health

Allan Shipp, M.P.H.
Deputy Director for Outreach
Office of Biotechnology Activities
Office of Science Policy
National Institutes of Health

Evadne Hammett
Project Officer
Clinical Research Policy Analysis
and Coordination (CRpac) Program
Office of Science Policy
National Institutes of Health

Liza Dawson, Ph.D.
Health Science Policy Analyst
Clinical Research Policy Analysis
and Coordination (CRpac) Program
Office of Science Policy
National Institutes of Health

Kimberly Allen
Program Assistant
Office of Biotechnology Activities
Office of Science Policy
National Institutes of Health

Michelle L. Saylor
Program Assistant
Office of Biotechnology Activities
Office of Science Policy
National Institutes of Health

Tara Hurd
Program Assistant
Clinical Research Policy Analysis
and Coordination (CRpac) Program
Office of Science Policy
National Institutes of Health