FOOD AND DRUG ADMINISTRATION (FDA)

AMERICAN SOCIETY OF CLINICAL ONCOLOGY (ASCO)

MINIMAL RESIDUAL DISEASE (MRD) AS A SURROGATE ENDPOINT IN ACUTE LYMPHOBLASTIC LEUKEMIA (ALL) WORKSHOP

Wednesday, April 18, 2012

8:00 a.m. to 4:00 p.m.

FDA White Oak Campus

White Oak Conference Center

Building 31, The Great Room

10903 New Hampshire Avenue

Silver Spring, Maryland

Meeting Roster

PANELISTS

Adam Bagg, M.D.
Professor of Pathology and Laboratory Medicine
Director, Hematology
Medical Director, Clinical Cancer Cytogenetics
University of Pennsylvania

Andrea Biondi, M.D.
Professor of Pediatrics at the Faculty of
Medicine and Surgery
University of Milano-Bicocca
Director of the Department and the School of
Pediatrics
University of Milano-Bicocca
San Gerardo Hospital, Monza, Italy
Head of the M. Tettamanti Research Center and S.
Verri Cellular and Gene Therapy Laboratory,
San Gerardo Hospital, Monza, Italy

PANELISTS (cont.)

Rita Braziel, M.D.
Medical Director, Hematopathology
Oregon Health and Science University

Giovanni Cazzaniga, Ph.D.
Senior Scientist
M. Tettamanti Research Center
Department of Pediatrics
University of Milano-Bicocca
San Gerardo Hospital, Monza, Italy

Meenkashi Devidas, Ph.D.
Research Associate Professor
Department of Biostatistics
Co-Director, Children's Oncology Group Statistics and Data Center, University of Florida site
Lead Statistician, Children's Oncology Group Acute Lymphoblastic Leukemia Disease Committee
Associate Program Director, Children's Oncology Group

PANELISTS (cont.)

Maria Grazia Valsecchi, Ph.D.
Professor Medical Statistics Centre of Biostatistics for Clinical Epidemiology Department of Clinical Medicine and Prevention University of Milano-Bicocca, Milan, Italy

Julie M.Gastier-Foster, Ph.D.
Director, Cytogenetics/Molecular Genetics Laboratory Associate Professor-Clinical College of Medicine
Ohio State University
Nationwide Children’s Hospital, Columbus Ohio

Nicola Gökbuget, M.D.
Head of Study Center, University Cancer Center
Goethe University Hospital
Department of Medicine II
Hematology/Oncology
Frankfurt, Germany

PANELISTS (cont.)

Stephen Hunger, M.D.- Co-Chair
Ergen Family Chair in Pediatric Cancer
Director, Center for Cancer and Blood Disorders
Children's Hospital Colorado
Chief, Section of Pediatric
Hematology/Oncology/Bone Marrow Transplantation
Professor of Pediatrics
Univ. of Colorado School of Medicine

Michael Link
President, American Society of Clinical Oncology
Professor in Pediatric Cancer, Stanford University

Wendy Stock, M.D.
Professor of Medicine
University of Chicago Comprehensive Cancer Center

Jerald P. Radich, M.D.
Associate Professor, Medial Oncology Division
University of Washington School of Medicine

PANELISTS (cont.)

Elizabeth Raetz, M.D.
Associate Professor; Fellowship Dir of Pediatric Hematology/Oncology
Department of Pediatrics (Oncology Division)
NYU Pediatric Hematology Oncology

Jacques J.M. van Dongen, M.D., Ph.D.
Department of Immunology
Erasmus Medical Center, University Medical Center Rotterdam, the Netherlands

James Whitlock, M.D.
Division Head, Haematology/Oncology
Women's Auxiliary Millennium Chair in Haematology/Oncology
Director, Garron Family Cancer Centre
Senior Associate Scientist, Child Health Evaluative Sciences Program, Research Institute
The Hospital for Sick Children
Professor of Paediatrics, University of Toronto

PANELISTS (cont.)

Brent Wood, M.D., Ph.D.
Department of Laboratory Medicine
University of Washington

FDA

Richard Pazdur, M.D.
Director
Office of Hematology and Oncology Products (OHOP)
Office of New Drugs, CDER, FDA

Gregory Reaman, M.D.- Co Chair
Associate Director
Office of Hematology and Oncology Products (OHOP)
Office of New Drugs, CDER, FDA

Albert Deisseroth, M.D., Ph.D.
Medical Team Leader
Division of Hematology
Office of Hematology and Oncology Products (OHOP)
Office of New Drugs, CDER, FDA

FDA (cont.)

Elizabeth Mansfield, Ph.D.

Director, Personalized Medicine Staff

Office of *In vitro* Diagnostic Device Evaluation and Safety (OVI), CDRH, FDA

Mark Rothmann, Ph.D.

Office of Translational Science

Office of Biometrics/DBV

Office of New Drugs, CDER, FDA

Marc Walton, M.D., Ph.D.

Associate Director

Office of Translational Sciences

C O N T E N T S

AGENDA ITEM PAGE

Welcome and Workshop Objectives

Gregory Reaman, M.D. 12

Regulatory Background and Considerations

Clinical Benefit for New Drug Approvals

Albert Deisseroth, M.D., Ph.D. 20

In vitro Diagnostics - Role in Prognostic and Efficacy Biomarker Assessment

Elizabeth Mansfield, Ph.D. 26

Biomarker Qualification: the FDA Perspective

Marc Walton, M.D., Ph.D. 42

Clarifying Questions to the Presenters

Expanding the Role of MRD from Prognostic to Efficacy Biomarker: A Surrogate Predicting Clinical Benefit

MRD as a Prognostic Biomarker and Potential Surrogate in Select Populations

Adults with ALL: The GMALL Experience

Nicola Gokbuget, M.D. 63

C O N T E N T S (continued)

AGENDA ITEM PAGE

Relapsed ALL: Does End-Induction MRD Predict 1 yr or 2 yr EFS

The COG AALL01P1 Experience

Elizabeth Raetz, M.D. 79

MRD as a Surrogate Endpoint for Efficacy Evaluation of New Drugs in Very High Risk ALL

Prognostic Significance of MRD in Ph+ ALL: The COG Experience

Stephen Hunger, M.D. 91

MRD as a Prognostic Marker in VHR ALL

Andrea Biondi, M.D. 103

Statistical Considerations for Surrogate Endpoints

Can a Critical Threshold Value of Early MRD Assessment be a Surrogate for OS/EFS in ALL?

Mark Rothmann, Ph.D. 116

Maria Grazia Valsecchi, Ph.D. 122

Meenkashi Devidas, Ph.D. 130

Clarifying Questions 138

C O N T E N T S (continued)

AGENDA ITEM	PAGE
Technical Approaches to MRD Assessment: Sensitivity, Feasibility, and Comparability	
Technical, Economical, and Validation QC Considerations for Multicenter MRD Assessment as an Efficacy Endpoint in ALL	
Jacques J.M. van Dongen, M.D.	144
Flow Cytometry in MRD Assessment	
Brent Wood, M.D., Ph.D.	165
Molecular Quantification of MRD	
Gianni Cazzaniga, Ph.D.	200
Clarifying Questions to the Presenters and Group Discussion	212
Questions to Workshop Participants, Decisions and Summary	219
Questions to Workshop Participants, Decisions and Summary (cont.)	288
Adjournment	321

P R O C E E D I N G S

(8:03 a.m.)

Welcome and Workshop Objectives

DR. REAMAN: Good morning. It's 8:00 or a little bit after, so I think we'll try and get started. We have a pretty packed agenda this morning, and I would like to keep on time.

So on behalf of Steve Hunger and myself, I'm Greg Reaman from the Food and Drug Administration, the Office of Hematology and Oncology Products. We'd like to welcome you to this workshop on minimal residual disease and acute lymphoblastic leukemia. I'd also like to acknowledge the support and assistance from the American Society of Clinical Oncology and recognize its president, Dr. Michael Link, who helped in the planning and execution of this workshop.

So I think before getting started, I'd like to ask all of our speakers around the table if they would be willing to introduce themselves, and maybe starting with Dr. Whitlock over in the corner. And we'll take it from there.

Jim?

DR. WHITLOCK: Jim Whitlock, Hospital for Sick Children and vice chair of COG ALL Committee for Relapse.

DR. STOCK: I'm Wendy Stock. I'm from the University of Chicago and I'm the co-chair of the leukemia steering committee for CTOC.

DR. GASTIER-FOSTER: I'm Julie Gastier. I'm at Nationwide Children's Hospital in Columbus, Ohio, and I run the molecular reference lab for COG.

DR. BRAZIEL: Rita Braziel. I am professor in the Department of Pathology at Oregon Health Sciences University and director of hematopathology.

DR. ROTHMANN: Mark Rothmann. I'm a statistician at the USFDA.

DR. WALTON: Marc Walton. I am the associate director for translational medicine in the office of translational sciences in FDA.

DR. MANSFIELD: Liz Mansfield. I'm the director of the personalized medicine staff in the

office of in vitro diagnostic devices at FDA.

DR. DEISSEROTH: I'm Al Deisseroth. I'm with the Division of Hematology Products at the FDA.

DR. PAZDUR: I'm Richard Pazdur. I'm the director of the Office of Hematology Oncology Products.

DR. FARRELL: I'm Ann Farrell. I'm director of the Division of Hematology Products.

DR. HUNGER: Stephen Hunger from Children's Hospital in Colorado. I'm the chairman of the Children's Oncology Group ALL committee.

DR. LINK: I'm Michael Link. I'm from Stanford University, and I'm the president of ASCO this year.

DR. BAGG: Hi. I'm Adam Bagg. I'm a hematopathologist at the University of Pennsylvania.

DR. GOKBUGET: Hello. I'm Nicola Gokbuget from the Goethe University in Frankfurt, Germany, and I am coordinator of the German Adult ALL Study group.

DR. RAETZ: I am Elizabeth Raetz from New York University. I'm one of the vice chairs for the COG ALL disease committee for frontline clinical trials.

DR. BIONDI: I am Andrea Biondi from Milano, from the University of Milano. And I am heading the pediatric department. And our center is the coordinating center for the ALL study group in Italy. And we are the reference lab for the PCR MRD detection in such study.

DR. CAZZANIGA: I am Gianni Cazzaniga from Monza, Italy. And I am running the reference lab, as I said, by Andrea Biondi, for molecular diagnostics in child ALL in Italy.

DR. VALSECCHI: Maria Grazia Valsecchi, biostatistician, University of Milan. I've been involved in trials on ALL for AIEOP and I-BFMSG study group.

DR. DEVIDAS: Mini Devidas, statistician from the University of Florida and lead statistician for acute lymphoblastic leukemia trials for COG.

DR. VAN DONGEN: I'm Jacques van Dongen from Erasmus Medical School in Rotterdam, and I run the MRD diagnostics in the Netherlands. And I'm the chairman of UMRD and the chairman of EuroFlow.

DR. REAMAN: Thank you. And I'm really pleased that we could have such an international representation at this workshop on minimal residual disease.

As you all know, the concept of detection of subclinical numbers of leukemia cells and otherwise normal bone marrow now has a history of more than three decades, beginning with the observation in George Janossy's laboratory, where leukemia cells were detected with aberrant membrane receptors.

Obviously, there's been improved specificity and sensitivity that's been enabled by progress in understanding lymphoid and myeloid differentiation and the molecular biology of leukemogenesis, to the extent that MRD has emerged as the most important prognostic factor in childhood ALL, as evidenced by multiple clinical trials in Europe and North America, involving greater than 9,000 patients.

These data from the COG 9900 study demonstrate, in a multivariate analysis, the prognostic significance of end induction or day 29 marrow minimal residual disease levels of less than 0.01 percent, surpassing, in importance, all of those prognostic features and factors, which have been conventionally utilized also for decades.

From the same study, this represents graphically the critical threshold value of .01 percent in its predictive ability for five-year event-free survival in children with ALL. And the predictive ability of relapse extends to both early relapses or treatment failure prior to the cessation, electively, of maintenance, as well as late relapses occurring following the completion of maintenance.

So I believe that MRD unequivocally, from the data that have been collected from multiple reports, is an established prognostic marker in ALL and has, in fact, eclipsed or supplanted most conventional risk factors for stratification on treatment, risk-adjusted treatment, studies.

I think that is also classically evident here, or graphically evident here, from the Italian and BFM cooperative group study in ALL. And within each classically defined or conventionally defined risk group, MRD has a prognostic significance and has become an important factor in stratification for risk-adjusted protocols.

We have a very focused objective at this workshop, and it really relates to the potential role of MRD in the regulatory approval process of new drugs for the treatment of ALL in children and adults.

We'd like to accomplish some understanding in the process for qualification of early MRD as an efficacy response biomarker. And this will be done within the regulatory framework of the FDA and the Office of Translational Science within CDER.

We'd also like to develop a consensus position within that regulatory framework, also have an understanding of clinical benefit and the importance of clinical benefit, and the potential role of surrogates in determining clinical benefit.

In that light, we will also have some presentations on the statistical requirements that will be required to advance MRD as a surrogate or for consideration as a surrogate endpoint.

Most of the data that have been collected and published to date utilize two different methodologies and is one of these methodologies superior to the other; so if we can develop some consensus on what might be a preferred technology platform if both technology platforms are required, and if so, in what situations they may be required, and define specific performance characteristics.

Also, in light of a recent NCI workshop, develop a consensus on standardized methodology and the need for centralized performance of MRD determinations. And then finally, to determine the need for an FDA-approved in vitro diagnostic if we're going to use MRD as a surrogate endpoint for regulatory purposes.

So without further ado, I'd like to get started and have Dr. Al Deisseroth just give a background on clinical benefit considerations in

drug approval.

Presentation – Albert Deisseroth

DR. DEISSEROTH: Thanks, Greg.

So I'm going to briefly discuss the standards that the FDA -- regulatory standards for definition of clinical benefit and what surrogates can be used in the regulatory context; and then discuss two examples of the development process that led to the acceptance of measurement of subclinical or minimal residual disease post-therapy as regulatory endpoints at the FDA.

Now, the basis of new drug approval at the FDA rests on the demonstration of efficacy and acceptable safety by adequate and well-controlled clinical trials. And one of the goals of this process is to be able to generate information that will enable product labeling that defines an appropriate patient population and adequate information for the safe and effective use of the drug.

There are two types of approval at the FDA, regular approval, in which substantial evidence of

clinical benefit is provided before regulatory approval, and then accelerated approval, in which evidence pertaining to a surrogate of benefit is provided before approval, but the evidence documenting benefit is provided in a post-accelerated approval study.

Here are examples of the accelerated approval process. The surrogate endpoints are reasonably likely to predict benefit. The diseases must be serious or life-threatening. There must be meaningful therapeutic benefit to patients, or existing or available therapies. And as I said, a confirmatory trial that addresses a benefit endpoint is required.

The final slide on FDA regulatory context is regulatory flexibility. And this means that the FDA must exercise its scientific judgment to determine the kind and quantity of data and information of an applicant is required. And this is what creates a vibrant regulatory environment that is able to meet the challenge of new classes of drugs as they emerge.

So I'm going to talk about two examples at the FDA that involve the development of measurement of subclinical disease. The first is antiviral HIV drugs.

Now, in the beginning of the last decade of the 20th century, the endpoints that were required rested on proteins, CD4 level and p24 antigen level. And the patients were forced to watch their CD4 counts decline before a management decision was made in terms of change of therapy. So this was a situation that cried out for change.

Over the ensuing five years, a task force, composed of industry, academia, NIH, and FDA, worked together to review the results of over 5,000 HIV patients undergoing therapy to test for correlations between the existing regulatory endpoints and a surrogate for benefit viral load, as measured by RT-PCR. And this effort showed if the viral load, post-therapy, was reduced to less than 50 transcripts per milliliter of plasma, there was likely to be a durable response and a reduction of progression of death.

So in 1996, there was a regulatory advisory committee meeting -- 1997 -- that established two endpoints that involved minimal residual disease: accelerated approval, which was a reduction of the viral load for 24 weeks, and then regular approval for individuals or drugs that could suppress the viral load for 48 weeks. And currently, viral load is a surrogate that's used for regulatory decision making in the United States.

The second example pertains to one that's familiar to many of those in the audience that have contributed to the development of this process, and that is the measurement of transcripts encoding the Bcr-Abl fusion protein in white cells among patients who have achieved a complete set of genetic remission. So it's a measurement of subclinical disease.

This effort has involved regulatory agencies and task forces on both sides of the Atlantic Ocean. The European program standardized the RT-PCR in 2003. There was a consensus conference

1 held at the NIH in 2005.
2 The major vehicle clinically for examination
3 of RT-PCR for Bcr-Abl as a minimal residual disease
4 endpoint was the IRIS trial. And this was a study
5 of over 1100 patients newly diagnosed, randomized
6 to two therapies, one of which was a kinase
7 inhibitor, imatinib. And the primary endpoint of
8 this trial was progression-free survival, secondary
9 endpoints, complete cytogenetic remission.
10 But in this trial was an exploratory effort
11 to determine the prognostic value of measuring
12 RT-PCR for Bcr-Abl. And this required the
13 development of a standard assay of normalizing the
14 Bcr-Abl transcript levels by the transcripts for
15 the normal gene, and then comparing that level
16 after therapy with the pre-treatment level that was
17 derived from a panel of 30 patient samples that
18 were used to define pre-treatment chronic phase
19 levels that were carried out by three reference
20 laboratories. And the reduction goal was to
21 achieve a 3-log reduction of this ratio.
22 At one year, it was apparent that the

minimal residual disease major molecular remission and surrogate endpoint was tracking with the complete cytogenetic remission. And by seven years, the achievement of the major molecular remission, at 18 months, correlated with retention of complete cytogenetic remission. So it was an endpoint that was predictive of outcome.

Now, during the time that this trial was taking place, the IRIS trial, the same sponsor of the IRIS trial initiated a prospectively randomized trial to address the efficacy of a second generation kinase inhibitor, nilotinib. And in this particular trial, that started in 2006 and ended accrual in 2008, the primary endpoint was a major molecular response at 12 months.

So this was the first clinical testing in this particular disease at the FDA in which an endpoint measuring subclinical residual disease was utilized.

The data lock was in 2009. And this particular dataset was brought forward to the FDA for regulatory approval, requesting accelerated

approval for previously untreated patients. And as you can see by this slide, the major molecular remission at 12 months showed that the second generation had twice the rate of a 3-log reduction as the first in class. And so this particular trial, that involved a subclinical surrogate of benefit of minimal residual disease, led to a regulatory decision, approving, through the accelerated pathway, first-line patients with CML.

So what can we conclude? What observations have we made about the steps involved in the development of minimal residual disease endpoints as a regulatory endpoint?

The endpoint must be identified in clinical trials. An assay must be available. It must be standardized. And then the endpoint must be applied prospectively in a clinical trial. So this hopefully brief summary will serve as a context for the development of the discussion throughout the day. Thank you.

Presentation – Elizabeth Mansfield

DR. MANSFIELD: Good morning. As I

introduced myself before, I'm Liz Mansfield. I run the personalized medicine program in the Office of In Vitro Diagnostics. And I'd like to give some background that I hope will help inform our discussion on the testing that may be required for MRD determinations.

So I'm here today, and we're talking about leukemia drug development using a novel or abbreviated endpoint, in this case, MRD. And we'll call MRD a biomarker. And this biomarker would define an endpoint that occurs prior to a standard endpoint such as progression-free, or event-free survival, or overall survival.

The criterion for use of a biomarker is that that biomarker, such as MRD, is acceptable to the therapeutic review office or division. And there are several ways to achieve this, to get the biomarker accepted.

The first is that you have a one-time or a case-by-case agreement with the review division to accept the use of the biomarker as a valid stand-in for another type of marker. Another type that Marc

Walton will discuss in the next talk is a biomarker qualification through a programmatic approach, which allows general qualification of a biomarker, and, thus, its broader use in the same context across several different drug development programs. Under this condition, this would be a surrogate under the particular context of use where the biomarker was approved.

I think Marc will go through this a little bit more, but it's very important to understand in terms of biomarker qualification, that through the CDER process of biomarker qualification, the biomarker itself is qualified. But a particular test strategy, or protocol, or materials are not necessarily approved. It only means that the review division accepts the biomarker meaning as a surrogate for another type of measure.

In order to use a biomarker in drug development, a cleared or approved test, however, is not required. But we do want strong evidence of analytical performance of the intended system. That's an expectation. And Marc will talk about

the relationship of the measurements used in biomarker development and qualification.

Another very important point to understand is that a biomarker is not an in vitro diagnostic test. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal or abnormal biologic processes, or response to drug, or so on. And in qualifying a biomarker, we are assessing its suitability. We are not assessing its suitability for clinical management. We are assessing its ability to stand as a surrogate for another measurement.

An in vitro diagnostic product or a clinical laboratory test is a completely different animal. Those are regulated through the Center for Devices at FDA and includes reagents, instruments, and systems that are intended for use in the diagnosis of disease or other conditions and so on.

In the case of an in vitro diagnostic product, we are looking, through our pre-market review process, directly at the safety and effectiveness of that test for clinical management.

So there's a very big difference between the concept of a biomarker and the concept of an in vitro diagnostic test.

Yet another important inequality here that's critical to understand, especially as we at FDA have gained a better understanding of how testing can affect clinical trials and the use of a drug post-approval, is the CLIA or Clinical Laboratory Improvement Act. It is not the same as the FDA law, the Federal Food, Drug, and Cosmetic Act. They have completely different goals and they look at different things.

So CLIA is administered through the Centers for Medicare and Medicaid Services, called CMS. And they regulate clinical laboratories. They regulate the activities in the laboratory, the protocols that are used, the personnel that run the protocols, and run the lab, and so on. They don't regulate the tests that are performed.

The FDA, on the other hand, under its authority in the federal Food, Drug, and Cosmetic Act, and in the case of IVDs under the device

amendments, regulates medical device manufacturers and actually regulates the test articles that are used in testing. We don't regulate the laboratories. We regulate the articles that they use.

The Food, Drug, and Cosmetic Act affects both investigational testing, that is, tests that have not yet been cleared or approved and are used in investigations, as well as commercial testing, that is, cleared and approved tests that are used in clinical laboratories.

It's important to understand, if you have heard the issues about laboratory-developed tests and the fact that, by and large, they are offered under FDA's enforcement discretion, that means we don't require them to comply with our laws.

Many people don't understand that if you're using these LDTs in an investigational manner, although enforcement discretion is available for their use in clinical diagnosis, in an investigational sense, they are still subject to FDA authority. So in investigational use, we still

need to see them even if they're LDTs.

So what are we looking at for in vitro diagnostic tests when we do a pre-market review? For this slide also, an investigational status is analytical validation or how well the test actually measures the analyte or substance that it's intended to measure. And analytical validation can include many different types of parameters, some of the most common of which are accuracy, and that is that the measurements actually represent the intended analyte and that the measurements are not biased in any systematic or random way.

We look for reproducibility, that is under constant conditions. When you hold the conditions that you're testing, you get the same result over and over. And then when you systematically vary conditions, that you get the same result over and over. And this has an element of robustness to it, that the test is robust to slight variations and conditions.

We look to applicable standards such as those available from the Clinical Laboratory

Standards Institute, or CLSI. For direction, that is available to laboratories in how to carry out these measurements.

For clinical validation, this is where, when we're looking at pre-market review, we need to understand what that test is intended to be used for. We need a very-well specified intended use. And the intended use tells us what is being measured, in what population, and for what purpose, among other things.

So this has to be spelled out very carefully before the test is validated. And what we're looking for is a demonstration of safety and effectiveness for that particular intended use. And we want test developers to pay special attention to the intended-use population, that is, you need to study those people that the test would ultimately be used in; and, generally, special attention to cutoffs or cut points, that is, where the answer that you get changes the clinical management that might be carried out, for example, from a positive or a negative, or a quantitative

value at which you would change clinical management.

So what are the effects of using a biomarker or a clinical trial test of a novel endpoint? I think what we're looking for here today is to see if MRD can be used as a novel endpoint for drug approval. And we think that it has the advantages of providing shorter clinical trials and perhaps faster access to important drugs for patients. And those trials actually could be smaller, depending on how they're designed.

What we're looking for is a high concordance of the biomarker to the traditional endpoint, that is, it needs to be very strong so that you are actually coming to the right conclusion by using a shorter endpoint or something that doesn't actually represent the clinical -- is not the clinical event of interest, but is a shorter-term biomarker of that. And we really want a very convincing link of that biomarker across the context in which it's used. And that can either be provided, as I mentioned earlier, on a case-by-case basis or

through a biomarker qualification program.

We do know from prior experience and from logic that different test versions that are used across different clinical trial sites can affect the readout if these tests are not standardized or accurately calibrated. So what we would be looking for in terms of tests that are used for this surrogate endpoint are some type of standardization so that we can compare site to site, test to test, and so on. Of course, the test must be validated and well documented.

Some possibilities are that this could be a cleared or approved test, for which we all know the clinical performance as well as the analytical performance, or in a clinical trial, you could agree on a testing paradigm that includes all the validations and controls that you would need to establish comparable performance across sites.

Let me jump a tiny bit into Marc's area now -- he'll go into this more deeply -- to talk about biomarker qualification. As I mentioned earlier, this is a programmatic qualification of a

marker, not a test, that's handled through the Office of Translational Science. And that requires that somebody submit a regulatory package or submission, called a biomarker qualification, that establishes the clinical meaning of that biomarker under a specific context of use.

The qualification is agreed to or declined by a review team that's put together and also includes people from the review divisions that are affected. And once this biomarker is qualified, essentially any other entity could use it under the same context.

The important thing to understand here is that biomarker qualification has no effect at all on the regulatory status of an in vitro diagnostic test. So in your minds, please separate biomarker qualification from test clearance or approval.

Of course, another method for biomarker qualification -- and I've already said this twice -- is one-time use that's agreed to under an IND protocol. And in that case, the use of the biomarker would not be transferrable to other

1 situations without a new agreement put in place by
2 the review division.
3 So in vitro diagnostic products are, under
4 our laws, divided up into three types of uses, that
5 is research use, and that is pre-clinical, not
6 necessarily according to a validated protocol;
7 investigational use, which is probably important
8 here, in which the diagnostic is used in
9 investigation either of the diagnostic itself or a
10 therapeutic product; and then finally, diagnostic
11 use or use once the product is cleared or approved.
12 For MRD tests, to date we have not cleared
13 or approved any test for minimal residual disease
14 in any setting for clinical management. So we
15 believe that any use in a clinical trial would be
16 investigational. It may be exempt investigational,
17 it may be non-significant risk, or it may be
18 significant risk, depending on how the biomarker is
19 used.
20 I think Marc will talk about this a little
21 bit. The endpoints in the clinical trial for MRD
22 would be pharmacodynamics, that is, you're looking

for response. You're not doing anything prior to treatment of the patient. And this, in our minds, becomes very similar to prognostic use, and in our treatment of a test, under the clearance or approval process, might require the same types of data.

One problem that we are looking at now is the current technologies and markers that are being used in clinical practice, primarily as LDTs, and can vary quite a lot. We don't know how they compare to each other. We don't know, for standard use, is one easier to use than another. Does it require more technical expertise? Does it have more variability? And ultimately, can these be standardized? Do we need to try to go to a single type of measurement technology? Can we standardize across different medical technologies and so on? And these are questions that we'll address later today.

So I'll just run quickly through the investigational status that I mentioned before. Significant risk investigations, if you're using a

test investigationally, significant risk investigation occurs when the use of the device in an investigation presents the potential for significant risk of harm to the patient when used as proposed.

In an IVD world, we typically see risk of harm when a test result is incorrect. So what we're worried about here is what happens to the patient if they get a wrong result. In the case of a significant risk investigation, we do require an investigational device exemption and approval for use of that test. And we're looking for how the subject safety is protected, what kind of knowledge is going to be gained from the investigation, and usually, not always, alignment with plans for later development and FDA review of that test.

We also have a class called non-significant risk devices. And the definition of this is that they don't present significant risk to the patient. We don't require an FDA submission. However, all the other investigational requirements that are found in the regulations still apply.

So how do we know when a device is ready for use to guide accrual, treatment, or assess outcome in a clinical trial? We want to see a fully specified device for purposes of the trial. And that means you have a protocol, you have set reagents, instruments, and so on, and that you understand the analytical performance of that device, that is, how well it measures and what the variability and biases might be.

We're looking for pre-clinical or clinical information that serves to justify patients' exposure to this test that has not yet been cleared or approved and, of course, a well-posed question or hypothesis that is answered or tested by the trial.

So let me summarize quickly here. Some of the main points I went over is that there is a substantial difference in the approach for biomarker qualification and the approach for test clearance or approval for commercial distribution, as well as a difference in the approach for investigational use of a test in a clinical trial.

MRD measurements are currently not widely standardized in the United States and we have no cleared or approved tests. We believe that having cleared or approved tests could be helpful.

The use of MRD measurement as a surrogate is dependent on the context of use, as well as the test performance, and I suspect that the Center for Drugs would be looking at both. And it's important to understand that since we have cleared and approved none of these, investigational use would be in play, even if you were using this in trials for response.

If this test were to be used for standard clinical management once a drug were approved on the basis of MRD, we would probably -- I mean, that could be available as an LDT, but we would like to clear or approve such a test.

I point you to -- and it's sort of chopped up here. If you want more information about medical device regulation, there's a website here that actually answers a lot of questions. And thank you very much.

DR. REAMAN: Thank you very much, Liz.

Our next speaker is Dr. Marc Walton, the associate director of the Office of Translational Sciences at the Center for Drug Evaluation and Research.

Presentation – Marc Walton

DR. WALTON: Good morning. I'm going to take a few minutes to talk to you about some of how we think about biomarkers and surrogate endpoints, and to include some of our idea about biomarker qualification here in CDER.

So the first thing I want to spend a few minutes on is just language, the nomenclature for biomarker types. It's important to ensure that we all understand each other when we're speaking. And we have found that, at times, people will be talking about the same thing and using different words or using the same words but meaning something different. So I'll take a moment to explain how we are using some of these words here at FDA.

The language that we use is focused on the way we use the biomarkers in therapy development,

which of course may be a little bit different than how some of these biomarkers and terminologies are used in clinical practice.

So there are a number of different biomarker categories. One is the prognostic biomarker. This is a biomarker that is telling us about what the future course of the patient is going to be, absent any subsequent therapy that alters their clinical course. So obviously, if we were to administer an effective therapy to the patient after measuring the biomarker, the prognostication of that biomarker becomes invalidated.

A predictive biomarker is also measured prior to administering an intervention, but it's going to be telling us not about the patient's future course, but about whether or not that patient has a greater or lesser propensity to have a response to the therapy that we're about to administer, and it can be a beneficial or a harmful response.

There are pharmacodynamic biomarkers. These are the response indicator type of biomarkers.

These are all measured post-treatment with the therapy, although often they may be evaluated as the difference between a pre-treatment and post-treatment measurement. And this biomarker is telling us whether and sometimes how large a biological response has occurred in that particular patient to the therapy they received.

My emphasis here is on a biological response. It could be a biological response of any particular kind. Many times it is a very early cellular response to the intervention binding to whatever cell it reaches in that cell, having a very immediate biochemical response, and not necessarily well linked to any subsequent changes in physiology.

Now, because it's telling us about the biological response to a treatment, like a predictive biomarker, which obviously is going to be treatment specific because it was telling us whether or not the patient could respond to a particular treatment, pharmacodynamic biomarkers may also be treatment specific, that is, for some

treatments, they may be a good indicator of a biological response; other treatments may not be a suitable indicator of biological response. It depends a great deal on what the intervention is and where it fits into the pathophysiology of the disease.

A subset, and generally a very small subset, of pharmacodynamic biomarkers are the efficacy response biomarkers, and also may be called an efficacy surrogate biomarker and sometimes just surrogate endpoint. These are typically going to be a very small subset. And these are the ones where we are going to interpret the response that we see as implying a true, meaningful efficacy response.

For FDA, these are the biomarkers that we can use as the basis of a drug approval decision because it is telling us about what will happen to the patient in their meaningful clinical outcome at some later time point. Like any other pharmacodynamic biomarker, these may be treatment specific. It depends a great deal on what the

biomarker is measuring and where it fits into the pathophysiology of the disease.

So there are these different uses that we give different names to. And the different uses of the biomarker depends upon the biomarker's characteristics. A particular biomarker is not necessarily restricted to one of these categories, and it can have utility in more than one of these categories. It will depend upon what the characteristics of the biomarker are, whether or not it has utility in these different ways. It also depends upon what is the question that we're asking of the biomarker as to which category we want to put it into.

Connecting back to something that Ms. Mansfield just said about how they view the pharmacodynamic biomarkers as being actually related in the way they review them to prognostic biomarkers, pharmacodynamic biomarkers, or the particularly surrogate endpoints, since we're interested in those here today, if one is using a surrogate endpoint in a drug development study and

comparing between treatments -- between a placebo and, say, an active treatment, in an archetypical randomized controlled study -- we may be using that biomarker to tell us whether or not that drug works at all and whether or not we should approve it. And that's using it as a pharmacodynamic biomarker.

On the other hand, one could use that same biomarker in a different study that was testing a treatment strategy system, where, based upon the biomarker, the change in treatment for a patient is being selected. And there, we're looking at the biomarker more as a prognostic biomarker. It's not telling us whether or not a particular drug should be approved. It's going to guide us in what treatment to select in that strategy because we believe it is telling us what would happen to the patient without further treatment.

Use in medical practice, when treatments are selected based upon a pharmacodynamic biomarker or change in treatment, is also being used in this prognostic way rather than as the surrogate endpoint of a drug approval trial.

Now, FDA does have concerns about pharmacodynamic biomarkers, and so we are very cautious in how we apply them, and particularly so when we're talking about the major decision in a drug development program, that is, the approval decision.

If a pharmacodynamic biomarker is discordant with the clinical outcome, it can mislead us in several ways. It could falsely indicate the presence or the absence of a benefit. It could falsely lead us to select what dose or regimen or population to treat with the drug by misleading us to make the wrong choice of dose. And it could also just give us an inaccurate estimate of the size or the frequency of benefit from treatment with the drug and lead to mistakes in designing a subsequent clinical trial.

This can happen from a number of different pathways. One is the potential for alternate mechanisms of action from what we think it is doing in a complex pathophysiology process. And another is what the actual shape of the relationship

between the biomarker and the clinical outcome is. And I'll go into a little bit more detail about this.

So this is our idealized situation for the ideal surrogate -- the ideal biomarker that can serve as a surrogate endpoint. We have a single set of pathophysiologic processes occurring in sequence. Somewhere in the middle of that is the biomarker that we're going to use as our surrogate endpoint. And later, after that, is the actual clinical outcome of the patient. And we have a drug that will affect these processes at the P1 point.

In this case, we all expect that whatever the surrogate endpoint tells us is happening to the patient is what's going to happen to the clinical outcome. But biology is often not quite so simple and there may actually be a complex set of pathophysiologic processes involved in the disease, where there are branchings of the processes. And one has parallel processes occurring that all contribute to the clinical outcome, as illustrated

in this very simple diagram.

So if we have in fact the drug hitting at the P1 point in this, then it's certainly going to alter the surrogate endpoint that we measure. But will it affect the clinical outcome? Well, that depends a great deal on what the balance of influence between the pathway that goes through the surrogate endpoint is, as compared to the pathway that does not go through the surrogate endpoint.

If the pathway going through the surrogate endpoint really is the dominant pathway driving the disease processes, then the surrogate endpoint will be a good guide to what's going to happen to the patient. If, on the other hand, the process labeled P3 that does not go through the surrogate endpoint is the dominant process, then the surrogate endpoint is going to mislead us.

In natural history, of course, the surrogate endpoint may be very well correlated with the clinical outcome. It's only when we have a drug intervention that discriminates between these pathways that we can see the discordance between

the surrogate endpoint and the clinical outcome. And of course, getting back to what I said earlier about more extreme cases of misleading us, if the drug intervenes at the P3 point, then the surrogate is misleading us because it's silent, even about effective drugs.

Now, moving on to what I said earlier about the shape of the relationship, even if the surrogate is in the pathway to the clinical outcome, there's still a question of what is the quantitative relationship between the surrogate that we measure and the clinical outcome.

So considering the case here, where we don't have the multiple pathways to worry about, then while we idealize this as thinking often about that dashed diagonal line in the middle as being the relationship, a nice, linear relationship, biology is often non-linear. And where the curve of relationship is, its shape, and where it's located has a big impact.

If, for instance, we're looking at a treatment that does alter the biomarker, and it

alters it from the B2 level down to the B1 level, and we know that in general, low levels of biomarker seem to have been correlated with good clinical status and high levels of the biomarker are correlated with poor clinical status, then altering the biomarker from B2 to B1 seems like it ought to have a meaningful impact to us. But it depends which of these relationships we're on.

If it we're that straight line, then you see in the middle the amount of impact that we would get on that dashed line is a pretty reasonable amount of impact of benefit for the patient. If on the other hand, the red curve is the operative curve, then what we're really seeing is just a treatment that can alter a patient from having a very poor outcome into a moderately poor outcome, very different.

If it were the bright green line, it's altering the patient's outcome from a very good to a very, very good outcome and probably not the treatment that we're hoping it is. If it were that middle green line, then in fact this amount of

alteration in the biomarker is providing us an enormous amount of benefit.

So which of these curves we're on makes a great deal of difference. And this quantitative relationship is dependent upon which curve we're on and which part of the curve we're operating between.

For a biomarker used as simply a cutoff value, then where we pick that B1 level is really critical and relates to which curve we're on. If we're on that bright green curve, then picking that B1 level anywhere around where it's shown here works very well as a good surrogate endpoint. If it is on the middle green curve, then the B1 biomarker, exactly where we pick it and exactly what the precision of the test is, can have a great deal of influence about how good a biomarker it is.

So I'll switch now to talking for a few minutes to talking about how we come to accept biomarkers in drug development. And generally, we have a long history of accepting biomarkers on a case-by-case basis in an individual drug

development program within a specific IND, NDA, BLA, or a labeling update.

A drug sponsor approaches us and brings us their biomarker information and the data that supports its use in the way they wish to. And we evaluate it within that specific drug development program and determine whether or not that use is justified. That use is going to be driven by that specific drug developer's needs and whether or not they can justify it.

Now, as we gain extensive amounts of this case-by-case experience and it becomes widely known, and the whole scientific community has a chance to evaluate it, we will often come to accept the biomarker's use in a more general sense and sort of prospectively accept it for a variety of uses in drug development, based upon this accumulated scientific information. But because the scientific information has not been developed in a focused way for a particular use, it usually requires an extended time frame to arrive at this point.

So at FDA, we really are interested in helping to advance drug development in a variety of ways, one of which we see as being the use of biomarkers in drug development. So in the future, as we move to the future, certainly the ways we have used biomarkers in the past and come to accept them remain entirely valid.

But we also see two additional ways that we want to emphasize. One is the co-development of a drug and a test, where the test is actually expected to be more than a drug development tool. It's expected to be used in clinical practice and would appear in the drug labeling, and would be what we call a companion diagnostic. And we have a policy guidance that was issued last year about this.

Companion diagnostics is a situation where we would expect the drug and the test to be approved at the same time, the drug of course by the Center for Drugs and the test by the Center for Devices. If, however, the test is not going to be used in clinical practice and it is to be solely

used to guide us in the drug development period, then it's not companion diagnostic. It falls into this other category that Ms. Mansfield was talking about. And for that, we have established a biomarker qualification program here. It is an outgrowth of the Critical Path Initiative that was articulated a number of years ago, that is much more widespread, about ways that we can improve drug development.

So biomarker qualification is a formal conclusion, that within a careful and very precisely stated context of use, the biomarker has been shown to reliably support a specified manner of interpretation and application in drug development.

Again, I'm emphasizing that qualification relates to use in drug development, and particularly regulatory decisions, as being central to the purpose of CDER's qualification. And it is also for biomarkers that are expected to have use in more than one drug development program.

Now, the context of use is really a critical

part of qualification. Having a clearly stated context of use is essential.

As conceptual examples, a prognostic biomarker, drawing from other areas in multiple sclerosis, we often use baseline measurements of lesion activity in the brain as a prognostic biomarker to select people who we want to enroll into clinical trials because we think they will have a greater propensity to have an MS relapse during the course of the clinical trial, and thereby give us the opportunity to show that the drug works.

So a prognostic biomarker might be -- gadolinium MRI lesions are a prognostic biomarker for use in selecting patients for MS clinical trials of anti-inflammatory treatments.

A prognostic biomarker -- in cancer, of course, we're seeing a lot of prognostic biomarkers being proposed for targeted therapies where a particular protein on the surface of a cell, which might be the target of a treatment, one needs to have that protein present so that protein can act

as a prognostic biomarker, if it's present, for the potential to respond to this targeted therapy.

Pharmacodynamic biomarkers, of course, we have a long history of using them as endpoints in phase 2 studies to help in the development of treatments in selecting doses. So we might have a particular response biomarker that we say is a pharmacodynamic biomarker for use in selecting dose levels for further investigation in phase 3 studies.

So going back to a bit of terminology again, I'm sure that all of this sounds a lot like talking about validation of biomarkers to you. So what's the difference?

The difference really is, what we talk about qualification here is what we meant by validation when we talked about validation for the past 10 or 15 years. The difference is, by changing the term, it allows us to have a conversation of what do we mean by validation, because many people have different ideas of what's meant by validation.

For us, a critical component of this is the

1 context of use. Qualification requires us to state
2 what is the precise context of use that we're
3 intending to consider the biomarker for.
4 Also on terminology, many of you may be
5 familiar with the IOM report last year about
6 biomarkers in chronic disease. And there, they
7 also talked about qualification, but they are using
8 the term in a little different way as non-
9 equivalent to the way we use the term
10 "qualification."
11 So what is it that becomes qualified? It is
12 the biomarker that becomes qualified. The
13 biomarker is a measurement of some substance, or
14 analytes, or some other objective describable
15 characteristic of the patient. It is not the assay
16 method that we use. It is not the device we use
17 for the measurement. It is the concept of what we
18 are measuring.
19 Now, of course, the assay method is
20 essential to measuring the biomarker. And we
21 cannot ignore that when we qualify a biomarker.
22 When we qualify a biomarker, we would talk about

this biomarker is qualified for this purpose. And we know that the biomarker is adequately measured when using these particular devices and assays.

So that would be part of what we describe as being our known body of reliance on the biomarker, but it is not intended that those assays are the only ways that it can be measured. We do believe that any one biomarker can have multiple ways of measuring it. And if they are equally good, then they are all suitable for use in clinical trials in measuring the biomarker and applying it in the qualified context of use.

It is just the assays' and methods' performance characteristics that are important to determining whether or not it's suitable for use in the clinical trial. Any assay methods that can be shown to be equivalent to those that established its qualification would be perfectly acceptable to us.

As Liz Mansfield talked about, qualification and CDER clearance or approval of commercial testing devices are two entirely separate things.

CDRH clears or approves the commercial testing devices for clinical measurements. Their clearance of a device does not mean that the biomarker has been qualified for a particular context of use within drug development, nor does qualification mean that that device has been cleared or approved by CDRH for use in clinical management of patients. The two regulatory authorities have different approaches.

Moving quickly, I just want to make it clear what the qualification's place in therapeutic development is. Qualification is not required for use of biomarkers. As I said earlier, the case-by-case approach remains entirely acceptable, and qualification is just intended for biomarkers that can be used in multiple drug development programs. Consequently, public knowledge will be essential to that.

There are three major parts to qualification that are not really very important to talk about here, and I will just skip over that.

The last thing I want to mention is when a

biomarker is qualified, though, FDA will make a public statement about that on the website and issue that statement as an appendix to our qualification process guidance. The status of that statement is of guidance. It is a formal FDA guidance. It functions as sort of a disease-specific mini guidance and carries the same commitment from FDA to stand by what we say. Thank you.

DR. REAMAN: Thanks, Marc.

I think rather than opening up for questions at this point, we may hold onto those until after the presentations on some clinical information. And I'm going to ask Dr. Hunger to introduce those speakers.

DR. HUNGER: Thank you.

The focus of the second session will be expanding the role of MRD from a prognostic marker to an efficacy biomarker or a surrogate predicting clinical benefit. Now, the first speaker will speak on adults with ALL, the GMALL experience. And that will be Nicola Gokbuget from Frankfurt.

Presentation – Nicola Gokbuget

DR. GOKBUGET: Good morning again. I would like to thank the organizers for inviting me, particularly all our pediatric colleagues, and I am really honored to be part of this meeting here. I will give you a more practical and clinical viewpoint on an efficacy biomarker, which I have just learned, which MRD probably is.

Since I am the only one representing adult ALL study groups, I would like to start with a brief introduction in the study group itself and our treatment protocols, which is important for the understanding of our investigations.

Now, the German Multicenter Study Group has conducted seven consecutive trials since 1981, and the chemotherapy is a BFM-based approach. We also conduct trials for older patients and related diseases. Since the beginning, the group has reference diagnostics and biomaterial banking. We have a database with more than 5,000 documented patients.

Very important is also this is a nationwide

study with more than 120 hospitals, so it really represents a kind of population-based data of real-world ALL treatment.

With these protocols, we were fortunately able to improve the outcome of all adult ALL patients from 24 percent survival in the trial, 03 in '87 to now more than 50 percent survival. In our most recent trial, this is a trial with targeted therapies, intensive chemotherapies, stem cell transplantation, and an MRD-based risk stratification.

In these trials, MRD was a major factor. We performed it since 1999 to evaluate MRD as a prognostic factor for MRD-based treatment decisions, as a new definition of response for future MRD-based clinical trials, which is the topic which we have today, and also methodological development, which I will not discuss today because that's not my topic.

The MRD logistics in our study group are working quite well, despite the large number of centers. The hospitals send diagnostic material to

the reference lab, which is important because this is needed for establishing the test.

If the material is insufficient, they get information from the diagnostic labs, then they send follow-up materials. The MRD result goes to the hospital and to the study center, and we give an MRD-based treatment recommendation to the hospital.

Now, the treatment protocol is risk stratified. I have mentioned before the risk stratification is rather easy in B- because of ALL high white blood cell count, Pro B-ALL. In T it is early T and mature T ALL. The Ph-positive ALL, T411, and late achievement of cytologic TR are all unfavorable prognostic factors. If a patient has no risk factor, he goes to the standard risk group. If he has at least one risk factor, it is a high risk group, and Ph-positive ALL is a very high risk group.

High-risk and very high-risk patients, according to these conventional factors, are allocated to a transplant in first CR. And in

standard risk, the new modality is that if the patient has a molecular failure, doesn't achieve after the induction and consolidation treatment, this patient becomes high risk as well and is a candidate for a stem cell transplantation.

This is the treatment regimen. We have initial, in the beginning, induction phase 1 and 2, BFM-based, then the consolidation. And afterwards, standard-risk patients have intensive consolidation treatment, hydromethotrexate, asparaginase, re-induction treatment. And in high-risk and very-high-risk patients, there is a stem cell transplant.

In parallel, we have a number of MRD evaluations in bone marrow. And the most important time points for treatment decisions are day 71, which is after induction, and at week 16, which is after the first consolidation. And with these evaluations, we try to identify patients with a molecular failure to offer them the transplant option. And also, we identify patients with a molecular relapse, again to offer them the

transplant. And I will come to the definitions of both of these groups in a minute.

The MRD evaluation in our group is based on individual rearrangements of immunoglobulin and TCR rearrangement. I will speak today only about the Bcr-Abl negative ALL. And it's done in a central lab in Kiel, headed by Michael Kneba and Monika Bruggemann. The method is standardized according to the published criteria. I think Jacques will maybe discuss it later on. There is an external, intensive control by labs, taking place all over Europe.

We also use for our analysis the standardized terminology for MRD-based response evaluation. This means we analyze molecular response only in patients with cytologic CR. This is clear. And then if the patient is MRD negative with a sensitivity of 10 to the minus 4, we speak about a molecular CR. If the patient remains MRD positive, we speak about a molecular failure. And during post-remission treatment, if the patient who has a prior molecular CR becomes positive again, it

is a molecular relapse. So quite similar definitions compare to the conventional cytologic definitions.

Now, I would like to discuss several issues regarding the prognostic value of MRD. And the first point is the molecular response compared to the cytologic response.

In our study, we have achieved now a very high rate of complete hematologic remissions on the left side, 89 percent of the patients, and there is no longer a big difference between risk groups. All of the risk groups have a high cytologic remission rate. But the molecular CR is only achieved in 69 percent of these patients, and there are very significant differences between groups, for example, a significantly lower molecular CR rate in high-risk patients compared to standard risk.

So this slide shows that MRD allows us a refined response assessment and that it can also be a tool for evaluation of new treatments. And I would like to show you one example of what we did

in our trials.

This was the use of rituximab in 320 positive standard-risk ALL. In these patients, we have added rituximab to the chemotherapy backbone, and we have compared the rituximab-treated cohort to a historical cohort without rituximab. And there was no change in terms of cytologic remission, 93 percent in both groups, but the molecular CR rate was significantly higher in the rituximab-treated group. And this also results in a significantly better remission duration.

So it is possible to use MRD to assess, for example, modifications of standard induction therapy.

The next issue I want to discuss is a time point, the most relevant time point for MRD evaluation. This is a publication from my colleague, Monika Bruggemann, in 2006. And she showed that in a very early time point, patients who achieve a molecular CR already at day 11 have a very good prognosis. Unfortunately, this is only 12 percent of the adult patients. On the other

hand, at a later time point, patients with a molecular failure have a very poor prognosis. This is a quarter of the patients.

So this means, at an early time point, we are able to identify patients with a low relapse risk. At a later time point, we are able to identify those with a high relapse risk.

The time point of evaluation is also important for treatment decisions. This slide shows you how the molecular CR rates develop during the induction therapy. During induction phase 2, there is still an increase of molecular CR rates, but between induction, before and after consolidation 1, there is only little difference.

Although the treatment includes high-dose ARA-C, high-dose methotrexate, and several other drugs, it only induces 10 percent additional molecular CRs, so that it shows very strongly that we deal with a population of patients who are chemotherapy resistant. So it will not help these patients to continue the conventional chemotherapy.

The next question is which level of MRD is

important. I think that this is not so difficult to answer. This slide shows you four different cut points for MRD and always at week 16, after consolidation 1. And with increasing MRD levels, the time to relapse is getting shorter. But this is not difficult to understand.

For the decision making, which level we should select for treatment modifications, I think two questions are important. First, the prognostic impact: is the group negative enough to justify a treatment modification? And second, which time do we have to realize treatment intervention? For example, if the MRD level is very high, we don't have time to change anything, find a donor for transplant, or something like that. Therefore, we decided to select the cut point of 10 to the minus 4 for our trials.

The next issue is about risk groups. The interesting thing about MRD is that it is an effective prognostic marker in all conventional defined risk groups. You can see here on the left side, our standard-risk patients with a poor

outcome of molecular failure, patients regarding remission duration, and also regarding survival. And on the right side, the high-risk patients and the same picture.

There is no difference in terms of prognostic relevance between both groups. And therefore, we did further analysis of the prognostic value together for high- and standard-risk patients.

This slide shows you this combined analysis. And the most important thing is on the right side, because there you can see the results if patients with stem cell transplantation are excluded. On the upper side is the remission duration, showing that molecular failure patients nearly all relapse in the long run. Only 12 percent continue complete remission if they are not transplanted. And on the lower right side, you see the survival, 33 percent survival. And please keep in mind that these relapses occur in the patients, although the intensive chemotherapy is continued.

Also, it is important to note that these

patients with the molecular failure have a poor survival of 33 percent, and this is our new, very-high-risk group. This is the poorest group we have in our adult ALL trials.

Just a short remark regarding older patients, the same; we have fewer data in the elderly patients above 55 to 65 years. But again, nearly all patients relapse who don't have a molecular CR. And the problem with these patients is that they even don't have the option to get a stem cell transplantation. So this is a group of patients where it's a real medical need, how to treat them if they don't respond to conventional chemotherapy.

I think it's also very interesting that results are quite comparable for pediatric and adult patients. On the left side is a publication by Conter, showing the overall survival in MRD high-risk patients, and on the right side, our own data. The risk groups are defined a little bit differently, but the outcome is quite similar.

The major difference is that, unfortunately,

in the adult population, we have far more MRD high-risk patients, 24 percent compared to 6 percent in the pediatric cohort.

The last issue which I want to mention is regarding prognostic value as a molecular relapse. This is something special in our trial, I think, because we followed up the MRD during the first-line treatment for a long time. And therefore, we were able to identify these patients.

We are able to show that patients with a molecular relapse who are not transplanted have only a chance to stay in remission of 10 percent. The median time to hematologic relapse is only three months, so the same as in molecular failure. And again, also, the survival is very poor if the patients are not transplanted in first CR. So I think we can combined the molecular failure and molecular relapse patients.

One question which we always ask is whether it is a worse effort to do all these MRD follow-up testings. Just to clarify the question, we compared the overall survival of patients with a

molecular relapse compared to those with a hematologic relapse. And the slide shows you what everybody can easily imagine, that patients with a molecular relapse, if they are treated, have a better outcome than if you wait for a full relapse.

This is very clear because, of course, in the full relapse, you have a higher tumor load. The patients are in poorer condition. They have complications. And in molecular relapse status, you have also more time to prepare the stem cell transplantation.

Now, finally, a few words regarding the clinical application of MRD testing. In our experience, there are several options to use MRD for treatment decisions. The first would be in molecular CR patients who de-escalate the treatment. This is something which is not an issue so far in adult ALL trials because we still need to improve our results. Our focus is more on the molecular failure patients and the attempt to intensify the treatment.

The only idea, unfortunately, which we have,

is the stem cell transplantation. And the question is whether this is a good treatment approach.

There are two problems with the stem cell transplantation. On the left side, you can see that it can only be realized in molecular failure patients in 38 percent of the patients. And the reason is simply that these patients develop very rapidly a relapse, and you cannot organize the transplantation during this time period.

On the right side, you can see that the patients who enter the transplant with molecular failure also have a poorer outcome after the transplant. This has also been shown by others, for example, by Dr. Bader from Frankfurt, for pediatric ALL.

So transplantation is probably not the ideal treatment in the patient with molecular failure. And we need some targeted therapies which work in these patients with chemotherapy resistance in order to avoid the relapse and in order to reduce the MRD level before the transplantation. And this is the most important question, which we would like

to discuss with regulatory authority.

Also, after the transplantation, we have options to modify treatment according to MRD, I want to mention as well. For example, it's already done that we modify immunologic treatment. We give DLIs or try to use experimental drugs in these patients because a full relapse after transplant cannot be cured anymore, at least in adults.

Finally, I would give you my opinion on MRD as an endpoint for clinical trials in ALL. I think it is a good endpoint because it is standardized, quantifiable, and reproducible. CR is already an accepted endpoint. MRD is nothing more than a better evaluation of CR.

The prognostic relevance is documented. The clinical benefit of MRD response is documented. We have historical reference groups available. Currently, we prepare European meta-analyses in adult ALL.

It has advantages for the trial design because you have early assessment of the response and less confounding factors. And I think this is

the most important thing, because outcome is depending strongly on stem cell transplantation, particularly in relapse patients. And this is a strong confounder for outcome.

Lastly, the examples are already ongoing for the blinatumomab trial, which takes place in Europe. But another point, MRD is also an indication for experimental treatments, and maybe we should discuss this as well, because already failure and relapse defined cytologically is an accepted entry criteria for experimental trials. And again, MRD is nothing else than a better method to detect these patients.

The molecular failure and relapse leads to cytologic relapses. We have a specific need in these patients because they are resistant to chemotherapy. And I think what Dr. Deisseroth mentioned before for the HIV patients, it is unethical. And for a clinician, it's really very hard to wait for a cytologic relapse and then to enter the patient in an experimental trial if you know already that the patient will relapse.

There are already approaches used, at least in our investigator-initiated trials. For example, we use nelarabine in MRD-positive ALL and also imatinib after transplantation.

Finally, I want to again acknowledge our study groups, and thank you for your attention.

I don't know whether we have questions or later.

DR. REAMAN: I think we'll wait until after all the presentations.

DR. HUNGER: Thank you very much. The next presentation will be from Elizabeth Raetz, who will discuss the Children's Oncology Group AALL01P1 experience in relapsed ALL.

Presentation – Elizabeth Raetz

DR. RAETZ: I'd like to start by thanking Dr. Reaman and Dr. Hunger, and the other organizers for the chance to review our experience in the pediatric relapsed ALL trial.

So the strategy that was taken by the Children's Oncology Group for relapse trial development was first to establish a safe and

tolerable induction platform regimen, and then later to use that platform to integrate new agents, and to use minimal residual disease response and complete remission rates at the completion of the first block of therapy as a method to screen for the activity and to potentially prioritize new agents for further study.

To that end, the COG ALL01P2 trial was conducted from 2003 to 2005. And this really established the safety and tolerability of this re-induction platform. The first new agent, then, to be combined with this platform was epratuzumab in the ADVL04P2 trial, which completed its accrual in January of 2011. And here in this trial, what it sought to determine was whether the remission re-induction rates were improved compared to historical control and whether a more favorable MRD kinetic pattern was observed with the addition of the new agent.

The objections of the O1P2 trial, the primary aim was really to develop a safe and active chemotherapy re-induction platform regiment for

relapsed ALL, which, as I mentioned, could be used to evaluate novel agents in future trials. And children and young adults, ages 1 to 21 years, with an initial ALL relapse involving the marrow, with or without concomitant extramedullary disease, were eligible for the trial.

In terms of timing of relapse, the majority, about two-thirds, of patients had an early marrow relapse, defined as occurring less than 36 months from diagnosis, 55 late-marrow relapses. There were 4 Ph-positive ALL patients who accrued. The vast majority of patients had a B lymphoblastic immunophenotype. There were only 7 TLL patients. The site of disease was 110 patients had isolated marrow or marrow in a non-CNS extramedullary site. There were 14 patients with concomitant marrow and CNS disease.

The study design for the trial is shown here. It was a phase 2 trial. There was an upfront randomization in terms of the order in which the blocks of therapy were delivered. Minimal residual disease was performed using flow

cytometry-based techniques in Michael Borowitz's lab at Johns Hopkins. It was performed after each of the blocks of chemotherapy.

The exception to the randomization was the CNS-positive patients. They were non-randomly assigned to ARM B so that they could have earlier introduction of high-dose ARA-C. This was a protocol exclusively for three blocks of re-induction therapy. Continuation therapy with either ongoing chemotherapy or stem cell transplant was at the discretion of the treating physician, depending on the site of relapse, timing of relapse, and donor availability.

This trial accrued from 2003 to 2005, and there were 124 eligible patients who enrolled. There was an initial cohort of 21 patients that received protocol therapy with idarubicin, but because that was too toxic, the regimen was modified to include doxorubicin during induction. And 124 patients who accrued subsequently to that will be the source of the data that I'll show.

The remission re-induction rates, we

published previously. If you look at all cohorts, the entire cohort of patients, and looking at those with a very early relapse, less than 18 months from diagnosis, their remission re-induction rate was 45 percent. The entire early cohort, less than 36 months from diagnosis, was 68 percent. And the late marrow relapse patients had a good CR2 rate of 96 percent.

In a separate analysis that included only B lineage, CNS negative, and Ph-negative patients, the remission re-induction rates were similar according to timing.

Minimal residual disease response rates according to timing of relapse are shown here. Again, this is MRD that was done by flow cytometry in Mike Borowitz's lab. It was done at the completion of each of three treatment blocks. And these data are for only those patients who achieved a morphological CR.

In looking at the left, in the early relapse patients, 75 percent of patients achieving a morphological CR were MRD positive. The positive

cut-point was just detectable MRD, so with flow-based assays, that was a sensitivity of about .01 percent.

If you look after block 2, 70 percent still MRD positive. And if you look after block 3, of those patients in remission, still, 43 percent were MRD positive. In looking separately at the late marrow relapse patients, 51, 41, and 25 percent were MRD positive at the completion of each of the three treatment blocks.

So I think we were all struck by the number of patients who, despite being in a morphological remission, had persistent minimal residual disease.

The EFS and overall survival have been updated recently by Xiaomin Lu, the statistician for this study. If you look at the whole entire cohort, at the B lymphoblastic patients, the five-year EFS is 29 percent, the five-year overall survival 40 percent, which compares very similarly to what historical regimens had shown.

If you look at the outcome now according to timing of relapse, again, exclusively in the B

lymphoblastic patients, if you look at those with recurrences of their disease within 18 months of diagnosis, their five-year EFS was dismal, 11 percent. If you look at the intermediate group, 18 to 36 months, their EFS was 24 percent. And if you look at those individuals with relapses greater than or equal to 36 months from diagnosis, their five-year EFS was 40 percent.

We next looked at the outcomes according to block 1 MRD response. So these are patients who were in morphological complete remission at the end of the first block. The MRD cutoff, as mentioned before, with the flow-based assay was 0.01 percent. If you look at all patients here, those who were MRD negative had a five-year EFS of 56 percent, compared to 15 percent in those who were MRD positive.

The analysis was done separately for those with early marrow relapse and late marrow relapse, and you can see there, the standard errors are high due to a small patient number. But again, the same pattern held true and was significant. The

five-year EFS for early relapse patients who were MRD negative at the end of the first block was 56 percent, compared to 8 percent in those who were MRD positive. And similarly, in looking at those with late marrow relapse, 57 percent for the MRD negative compared to a little bit better for the MRD positive, 24 percent five-year EFS.

The kinetic pattern of MRD response was also analyzed by Mike Borowitz on this trial. There were four general patterns of kinetic response that were seen. In the R1 group, those are patients that after the three time points, or the three blocks of chemotherapy, there were those patients who were negative at all three time points.

A second group were those who began positive and became negative with the subsequent two blocks of therapy. There were others who were positive and remained positive, but had greater than a 1-log reduction in their disease burden. And the fourth, smallest subset of patients were those that actually had rising MRD, despite being in remission.

As you might suspect, the kinetic pattern of MRD also correlated with response. The most favorable responses were seen with those who were MRD negative at all time points, the gold line at the top. And the worst overall or worst event-free survival rates were observed in those with rising MRD, the red line.

Next, I wanted to speak briefly about the successor study. This trial used the agent epratuzumab, a CD22 monoclonal antibody in addition with this re-induction platform. And this was the COG ADVL04P2 study. The primary aim of this study was to determine if epratuzumab, combined with the O1P2 backbone regimen, improved the rate of second remission in children and young adults with early marrow relapse.

Note, this was a study exclusively for the early marrow relapse population. Children and young adults, 2 to 30 years of age with first early marrow relapse, CD22 positive, with or without extramedullary disease were eligible.

The study design was shown here. It was a

two-part study. Part A was the phase 1 portion of this study, which accrued 15 patients. It was a feasibility design. Here, epratuzumab was given twice weekly as a single agent during the re-induction phase.

Response was assessed. And regardless of that response, individuals went on to receive four weekly doses in combination with the O1P2 platform regimen. Patients had their rate of second remission and minimal residual disease determined at the completion of the first block.

The second part of this study was a phase 2 pilot group-wide portion of the study. And here, epratuzumab initially was designed to be administered in four weekly doses, in combination with the triple induction platform. The endpoint of the study was CR2, looking for a 13 percent improvement in remission re-induction rate. Secondary endpoint was the MRD response rate.

There was an initial B1 cohort of patients that received epratuzumab on this dosing schedule. Then, while this was in progress, we had

preliminary PK data to suggest that the antibody half-life was much shorter in leukemic patients than it had been in adults with indolent lymphoma. So we accrued a second cohort, where epratuzumab was actually administered twice weekly for eight doses in combination with block 1 chemotherapy.

The feasibility portion accrued 15 patients, the B1 cohort, 56 patients, the B2 cohort, which included epratuzumab twice weekly for eight doses, accrued 16 patients, the last of whom enrolled in January of 2011.

The responses to the reduction phase in block 1 for part A of the study were previously reported, responses to antibody alone, 8 patients with stable disease, and then 4 with minimal remission cytolytic responses, 3 with progressive disease.

At the end of block 1, among the 15 patients, 9 achieved a complete remission. And notably, 7 of those 9 patients had no detectable minimal residual disease, which compared quite favorably to the regimen without antibody. Other

patients had stable disease, progressive disease, or died from infection, or removed from protocol therapy.

The responses for the B1 and B2 cohorts were recently analyzed. We reported these data at ASH this year. If you look at these cohorts of patients, among the 48 evaluable B1 patients and 50 B2 cohort patients, the remission re-induction rates were nearly identical to that observed with historical chemotherapy platform alone.

But in looking at those patients who achieved remission and had MRD responses assessable, we noticed a much more favorable pattern of MRD response, I think, similar, Nicola, to what you showed for your rituximab study. And here, 42 percent of patients were MRD negative, compared to 25 percent MRD negative, with early relapse with the platform alone.

So in conclusion, rates of MRD positivity are much higher in relapsed ALL than newly-diagnosed ALL, where about 25 percent of patients are MRD positive. Early MRD response was a very

strong predictor for event-free survival in the O1P2 pediatric relapse trial.

The kinetic pattern of MRD was also predictive of longer-term outcomes in relapsed ALL. And the combination of MRD response and timing of relapse identified different risk groups, suggesting that MRD may be helpful in stratifying salvaged therapy.

I'll just end by acknowledging my many colleagues who I have the pleasure and privilege of working with in the Children's Oncology Group. Thank you for your attention.

DR. REAMAN: Thanks, Elizabeth.

Our next speaker is Dr. Hunger, who's going to talk about the prognostic significance of MRD in a very-high-risk group of patients, those that are Ph-positive ALL in the COG experience.

Presentation – Stephen Hunger

DR. HUNGER: Thank you very much.

So these data are very similar to those that have been shown earlier, simply looking at the prognostic value of the minimal residual disease in

the most recent Children's Oncology Group high-risk ALL study and showing an ordinal response rate and outcome rate.

I would say, listening and sitting here, sometimes you feel like Alice in the looking glass, given that in every patient treated on clinical trials in North America and Western Europe, MRD is used as a prognostic factor and is used to allocate their therapy. And that is routine clinical practice.

So I think we need to somehow harmonize what's happened over the past decade and what the approval processes are in terms of federal regulatory agencies.

So I think a second and very different question is whether early response minimal residual disease can be used to assess the benefit of interventions. And I think we have to at least consider the fact that this may depend on the trial design. Are we comparing treatment A versus treatment B, treatment A versus treatment A plus drug X, treatment A plus drug X versus treatment A

plus drug Y? Those may have different implications for minimal residual disease assessment.

I would say that the available data on the A versus B design is, frankly, not encouraging in some settings. For example, the trial I just mentioned to you, we randomized patients to receive, in the context of a multi-agent chemotherapy regimen, 14 days of dexamethasone during induction, and 28 days of prednisone during induction, with 28 days of prednisone being the standard. And what we found was that there was an event-free survival advantage for 14 days of dexamethasone, but there was no MRD difference between those.

Now, it is possible, because MRD was measured at day 29, two weeks after the last dose of dexamethasone, that that might explain it. But I do think we have to acknowledge that the trial design may have a significant implication on how MRD is used to assess the efficacy of an intervention. And there are other examples where there is an event-free or overall survival

advantage, but not an early MRD reduction advantage.

So I want to use pediatric Philadelphia chromosome-positive ALL as an example here and just quickly go through a little bit of background on this for people who are not familiar with it. About 3 percent of childhood ALLs are Philadelphia chromosome-positive, and the rate begins to increase significantly in the late teens and early 20s. And this is a much higher percentage of adults, particularly older adults, with ALL.

Just like other types of ALL, age, white count, and early response are important prognostic factors in Philadelphia chromosome-positive ALL. And prior to the advent of imatinib, there was really a dismal outcome for this subset of ALL. And an international consortium, referred to as the Ponte di Legno group, that includes many of the people in this room, showed an initial publication that matched sibling transplant was better than chemotherapy, and then a second publication, both of whom by Maurizio Arico with the first author,

that any stem cell transplant was better than chemotherapy.

Then this slide summarizes some of those data, and I want to emphasize that these patients were all treated without tyrosine kinase inhibitors. So what we see in the pre-imatinib era is that the CR rate was 82 to 89 percent. In other subtypes of childhood ALL, it is typically 98 to 99 percent. And we see between the first and second era, modest improvements in seven-year event-free and overall survival that are statistically significant, although the clinical significance of some of those differences is modest at best.

So the conclusion I would make from this is, over time -- and the main difference in this time was more intensive standard chemotherapy and more widespread and earlier use of stem cell transplantation -- we saw modest improvements in outcome for patients with Philadelphia chromosome-positive ALL. But overall, the outcome was quite dismal, with less than 50 percent alive at seven

years.

After the advent of imatinib, a number of groups tried to understand how to incorporate imatinib into intensive chemotherapy regimens, the Children's Oncology Group and the ALL0031 study, led by Kirk Schultz -- and these data are published -- used a design where there was no imatinib given in the first month of therapy.

Patients then entered the study at the end of the first month, either in complete remission or not in complete remission, received two blocks of chemotherapy, and then were allocated to stem cell transplant of an HLA-matched-related donor, or to continued chemotherapy plus imatinib if they did not have a related donor.

This study was conducted as a dose-escalation study. And I'm really going to focus my discussion on the patients who received imatinib continuously, which we refer to as cohort 5. So these patients started imatinib at the end of induction therapy and received the drug continuously through the end of treatment, which

was approximately two and a half years following diagnosis.

This was a non-randomized historical control trial, so we see that the outcome of patients in cohort 5 enrolled in this study. And the top yellow line was much better than the outcome of historical controls, with three-year event-free survival rates of 80 percent versus 35 percent, so not subtle differences.

Then when we looked at the patients treated in cohort 5, although the numbers are small in these groups, very small, we saw no difference between those treated with chemotherapy plus imatinib versus those who underwent related or unrelated donor stem cell transplantation.

So the implications of this, at least to us, was that there was outstanding early outcome of intensive chemotherapy plus imatinib in Philadelphia chromosome-positive ALL. The follow-up, at least when this study was reported, was relatively short, but the excellent outcomes for cohort 5 patients continued.

All patients completed therapy by early 2009, so we're now more than three years after the last patient continued therapy, and we've not seen significant numbers of events following cessation of tyrosine kinase inhibitor therapy. A manuscript is being prepared to update those results.

There was no evidence that stem cell transplant was superior to chemotherapy plus TKI therapy. And it suggested to us that further improvements would come through optimizing TKI therapy, not through optimizing the chemotherapy regimen itself or the stem cell transplant regimen.

So we elected to pursue a successor trial that tested dasatinib, a more potent ABL-class tyrosine kinase inhibitor in a trial, AALL0622. And I want to focus a little bit on early results of that trial, simply as an example, so I want to contrast the TKI usage.

So the chemotherapy between the two trials is essentially identical. In the first trial, AALL0031, imatinib started at day 1 of consolidation, so after completion of four weeks of

induction therapy without TKI.

The successor trial used dasatinib, and it started at day 15 of induction therapy. Because this was a safety pilot, the first 41 patients enrolled received discontinuous therapy with dasatinib, meaning they received it for two weeks out of every three- to four-week chemotherapy block. When that was established to be safe, the following patients received continuous treatment with dasatinib, starting at day 15 of induction therapy.

So this design, although it is, again, a historical control retrospective design, allows us to at least compare the effect of dasatinib added to chemotherapy induction and also the effect of dasatinib versus imatinib on later response. And this is somewhat confounded by the fact that two-thirds of the patients enrolled in the dasatinib trial were treated discontinuously, so they had less exposure to dasatinib.

So the first question is, if you think, on the 0031 trial, patients received four weeks of

chemotherapy, then entered the trial, and on the successor trial, they received the same four weeks of the same chemotherapy, but the last two weeks of that, day 15 through 29, they received dasatinib.

So that gives us an ability to at least ask what were the complete remission rates with the addition of dasatinib, and what were the MRD rates with the addition of dasatinib.

So when we look at this with complete remission rates, or day 29 M1 marrow status, which is less than 5 percent blasts, we see on the 0031 trial, 89 percent of patients reached this criteria, which is more or less identical to historical controls. In contrast, in the trial where dasatinib was given for two weeks, now 98 percent of patients attained complete remission and that reached a statistical significance.

When we looked at the day 29 MRD response, this is done by flow cytometry by Michael Borowitz, who's sitting in the room, looking at patients denoting negative, is less than 10 to the minus 4. We see there are only 25 percent of patients

treated without TKI for four weeks achieve that level versus almost 60 percent of those who'd had two weeks of dasatinib added.

So our conclusion from these -- and these have been submitted for presentation. Bill Slayton is the study chair who conducted that trial. So our conclusions are that dasatinib significantly increased the remission rates at the end of induction and decreased the end induction MRD burden.

As I mentioned, these are historical control comparisons, but I think it's worth considering, what would the regulatory implications be if this had been a randomized trial of induction chemotherapy, plus or minus dasatinib. The results in the historical control comparison look quite striking.

So again, considering the design of this trial, we looked at this endpoint here, and now we're looking at the patients after two additional blocks of chemotherapy.

So all the patients treated on the first

study received this chemotherapy with continuous imatinib treatment for those approximately seven or eight weeks. The patients treated on the successor study received dasatinib for two weeks, and induction, and then received dasatinib for these two blocks of chemotherapy, although again, two-thirds of them received only half -- half the time were receiving dasatinib. A third of them all the time were receiving dasatinib.

So again, looking at minimal residual disease response, we see that 75 percent of patients treated with imatinib in that schema became MRD negative at the end of that second block of consolidation chemotherapy versus 89 percent of those treated with dasatinib becoming MRD negative, again reaching statistical significance.

So again, even using dasatinib discontinuously in the schedule, significantly increased the percentage of patients who became MRD negative, suggesting to us it's a better tyrosine kinase inhibitor than imatinib for Philadelphia chromosome-positive ALL.

Again, I think we have to consider what would the regulatory implications have been if this study was a randomized comparison of imatinib versus dasatinib. It was not, but what would they have been if you saw data such as this?

So I'm going to stop there. I want to thank and acknowledge the input of a number of people: Maurizio Arico, who is not here, in the Ponte di Legno group; Kirk Schultz, who is the chair of the Children's Oncology Group study that studied imatinib; Bill Slayton, who is the chair of the trial that studied dasatinib; Mini Devidas, sitting in the room, who's the statistician; and my colleagues, Bill Carroll, Naomi Winick, Elizabeth Raetz, and Mignon Loh. Thank you very much.

DR. REAMAN: Thanks, Steve.

I think we're actually scheduled for a break, but it may make more sense to complete these presentations. And so I'll ask Dr. Biondi to give us a presentation on the AIEOP-BFM experience.

Presentation – Andrea Biondi

DR. BIONDI: Good morning. Thank you very

much indeed for the opportunity of being here today and to share the experience that has been achieved in the last, I would say, more than 10 years within the AIEOP-BFM study group, and more recently, specifically, within the Italian and German study.

I would like to contribute to this workshop by trying to address the following question. Well, the first one is not a question, but just to reinforce what has been already said by previous speakers on the prognostic relevant impact of the MRD in risk stratification, and accordingly to tailor the treatment, but then to address three questions, why not correlation we've found so far with omics technology?

The second one, how to integrate MRD in the design of future ALL treatment protocols. And I will bring to you what has been the newest protocol within the German-Italian study group, chaired by Professor Schrappe, in order to highlight what is the perspective on using a designed clinical trial, the MRD. And finally, to discuss a few points on the MRD as a surrogate marker for a response

assessment in novel therapy.

It's been already stated by a previous speaker that today there is, at least from our perspective, no doubt that the MRD assessment is the best available method to assign the risk of ALL patients. Why? Because there is a close association between the quality of the molecular remission and the final outcome, independently of the applied treatments in childhood ALL.

I want to reinforce and highlight that it's still unknown why the exposure to drugs, during the early phase of treatment, which are not like the old drugs panel that the patient will use during the treatment, discloses different in vivo chemosensitivity, which influences the final treatment outcome.

So we should be aware that we are measuring something, but we still don't know the reason for such event.

These of course, in this first slide, I'm sure that Professor van Dongen further reinforced. And it's been already stated by FDA authorities

that it's an essential step. It was for us, but I'm sure that it should be for the future if we are planning to use this surrogate marker, the issue and the concept of standardization.

This just will give you how, for example, the application of the continuous definition, and not the single time point, is essential criteria, because if you are using molecular technology, you are getting number. And number is not the fixed, precise -- you are receiving an interval, which of course would require a definition.

These, within the AIEOP-BFM, has been strictly approached by Euro SG-MRD, and a guideline has been already defined and standardized. But I'm sure that Professor van Dongen will approach this issue even in more detail.

So this is just the results of the published data from The AIEOP-BFM, the Italian-German 2000 ALL study, in close to 5,000 patients. This is the final outcome in terms of survival in EFS.

When you look to the impact of the MRD stratification, based on the standard way to

classify the patient as standard risk or high risk according to NCI, you see that the three colored curves, which reflect the definition of standard-intermediate-, and high-risk patients, clearly define category of patients with a different outcome. And this is true according to immunophenotype.

In the TLL, we reported that the way the clearance occurs in TLL is different from BCP, from B cell precursor ALL. That should be kept in mind, of course, if you're using this in that context. But again, according to immunophenotype.

More interesting, even if you look to the assessment of early response -- which is still part of the BFM strategy, which is the first seven days of exposure to a single drug, an intrathecal methotrexate, which we call the pre-phase -- if you stratify the patient according to the response to that phase, which is still part of the process of stratification of the patient, you see that even within the so-called poor prednisone responder, you can identify a patient with a different outcome

according to the level of MRD. So no doubt that there is a close association between the MRD stratification of a patient and outcome.

This suggests a question to bring to your attention that might be worthwhile to discuss. Why no correlation we've found so far with omics profiling? I say that several attempts have been made to correlate MRD with gene expression profiling some leukemia cells or to germline or leukemia-associated gene polymorphing, just to investigate the individual variability to response.

We still don't know the reason for that. Probably because MRD is measuring a balance between these two factors, the heterogeneity of the leukemia and the heterogeneity of the host. And you are measuring something which is clearly individual based.

We don't know whether MRD could eventually be replaced by novel risk factors, unpresented features. But it's still unlikely to occur, because MRD measurements reflect the combination of leukemia-presented features as well, the efficacy

of therapy.

Now, we'll move to the next question, how to integrate. And just to give you a flavor of how a group dealing with a large patient population is moved in 2009 to a new protocol, which is chaired by Professor Martin Schrappe. So we have tried to, of course, address some randomized question. But again, I will give a sort of hint of what is the use, of how the use of MRD will help to refine patient stratification and, accordingly, to address in a better way randomized question to further improve the treatment of this disease?

I will give you two examples. The first one is based on what we achieve in the previous study. In what we define intermediate MRD study, we were able to look at the different time point, end of induction and end of the first course of consolidation, three months' time. And we clearly define that there is a population which clearly has a different time, different MRD levels, in the time point, and clearly depict in the lower curve of the diagram a population which clearly showed a worst

outcome, as compared to all the intermediate groups, according to the definition of MRD. And that is the reason why, in the new 2009, this patient will enter in the high-risk group, assuming that by giving a more intensive treatment, we will be able to achieve a better outcome. But again, this is just to give you a flavor of what is the idea of using MRD in a prospective in a new trial.

A second example, by following the different time points later on, what are planned by the protocol, our colleagues from Germany, more specifically Andre Schrauder and Gunnar Cario, reported that all patients at the end of three months were still positive greater than 10 to the minus 3, which is the definition of being high risk, according to MRD.

Those patients, during the progressive treatment, clearly display a clearance of the MRD levels. And of course, this is an essential pattern because it explains why in some of these patients, we did achieve a very good outcome despite being high risk.

But what we did learn is that there is a subgroup of patients who are still persistently positive at week 12, after the first three blocks, by waiting for the regenerating bone marrow. And these patients are now defined, based on the MRD, on continuous levels, as very high risk. And accordingly, it would be an ideal category to test an innovative strategy and to assess whether, in this category, having a better response or a better remission will have, as expected already, a better outcome if those patients will undergo a bone marrow transplantation.

We did learn, as many other groups, that being MRD positive before bone marrow transplantation, and more important, in those patients who are defined as very-high-risk patients, the chance of achieving EFS is rather different in these data from our single institution -- that has been still unpublished -- just to confirm what has been already presented by several study groups.

Accordingly, the question would be, is it

possible in these very high-risk patients to introduce an innovative therapy to reduce the MRD burden, and, accordingly, achieve a better remission before bone marrow transplantation and consequently having a better outcome? This is a second example on how to use MRD to refine the risk group assignment and address a clinical question.

Finally, let me just comment on what I hope will be a very intensive discussion later on. Can we use MRD as a surrogate marker? I would like just to share with you at least four thoughts that I believe are very important, at least based on our experience.

First of all, MRD measurement at different time points are likely to reflect a different biological phenomenon. Is the clearance of blasts versus the quality of remission, which is basically the way we are measuring in the protocol I've presented so far, a relevant issue?

I can show you just these data from, again, our study group. Gisette Basso from the Italian study reported the assessment of MRD by flow

cytometry at day 15. And according to color, you can clearly identify a patient with a different outcome.

If you ask the question, are those patients exactly the same as the patients that later on you've identified by using PCR in a different category? The answer is no. Assuming the clearance of blasts you're measuring at day 15, whether or not you're using the same method, it might be that at this time point, flow cytometry measuring valuable cells is better than PCR technology that can include even dead cell because they're measuring DNA.

This is just the first comment.

The second issue that we should be aware of, which can be another issue of complexity, is that for all childhood ALL, as has been shown, the clonal heterogeneity could be potentially another issue to be considered, or at least the term of pathophysiology is something relevant to be considered.

I give you an example in a cartoon, which

clearly shows how the color of the different type of rearrangements, which is a sort of proxy of the different clones, might change from time to time, suggesting the existence of this clonal heterogeneity.

Second comment. Time points for endpoints on activity, which means response, might not be the same as a surrogate endpoint of efficacy, survival, or event-free survival.

A third comment, which I think is very important, I personally believe that targeting therapy might have a better chance to be studied by using MRD as an endpoint of activity. And naturally, what we heard previously by Steve Hunger, when they reported their experience of this COG for dexa versus TKI in a Ph-Positive ALL, are probably aligned with this statement.

Of course, deciding the efficacy of treatment, which can be assessed in terms of molecular response, requires that MRD levels are properly validated, as we heard from an FDA authority previously.

So let me go to the last slide, just to say that there is no doubt today that MRD is generally used to guide post-induction as post-consolidation therapy. We'll see it is unlikely that MRD studies could be completely replaced by novel risk factors. The combined use, evaluation, and newly-available genomic information on leukemia-presented features will further improve risk assignment of ALL patients.

However -- and probably one of the reasons we are here today -- I would say more relevant for adults, but definitely relevant even for childhood ALL, clinically relevant improvements in ALL can only result if MRD-based stratification is paralleled by the finding of the appropriate therapeutic strategy. This is what our children need, and this is the reason why I thank you for your attention.

DR. REAMAN: Thank you, Andrea. We can take maybe 5 or 10 minutes for questions to the morning's presenters and then break, if there are any burning questions. Otherwise, we can do this

informally during the break.

Any questions from the panelists?

(No response.)

DR. REAMAN: If not, let's take a break and try and be back at 10:30. Thanks.

(Whereupon, a recess was taken.)

DR. REAMAN: Can I ask everyone to take their places? And we'll try and resume our program here.

So we're going to start this section of our workshop with a discussion on statistical considerations for surrogate endpoints. And leading off is Dr. Mark Rothmann from the Office of Biometrics at CDER in the FDA.

Dr. Rothmann?

Presentation – Mark Rothmann

DR. ROTHMANN: Hello. I am going to provide a general introduction on the topic of surrogate endpoints. I will briefly describe clinical benefit and surrogate endpoints, how a surrogate endpoint is used in the evaluation of a potential surrogate endpoint. I would also like to thank Dr.

Thomas Fleming for the use of some slides.

A clinical benefit endpoint is a direct measure of how a patient functions, feels, or survives. The NIH definition of a surrogate endpoint is a biomarker intended to substitute for a clinical endpoint.

Analysis on a surrogate endpoint can be used to draw conclusions about treatment effects on a clinical benefit endpoint, as in a regular approval, or can be used as an accelerated approval to determine whether the results on the surrogate endpoint reasonably likely predict clinical benefit. An important property of a surrogate endpoint -- for a biomarker to be a surrogate endpoint for a particular clinical benefit endpoint, effects of the intervention on the clinical benefit endpoint should be reliably predicted by effects on the surrogate endpoint.

The analysis of a surrogate endpoint needs to give you a guidance or impression of how the results will compare on the clinical benefit endpoint.

Evaluating whether a marker is a surrogate endpoint involves understanding of the disease process. Is the marker on the pathway towards clinical benefit? And it involves results from many randomized clinical studies that evaluate whether treatment effects on the surrogate endpoint correspond with treatment effects on the clinical benefit endpoint. This usually requires a sort of meta-analysis, and I'll talk a little bit more about that later.

There are a variety of ways a biomarker can be related to a clinical endpoint or a clinical outcome. There might not be a cause and effect relationship. The biomarker may be associated with clinical outcome, but not be on the causal pathway towards clinical outcome.

For example, CD4 counts are correlated with the likelihood of mother-to-child transmission of HIV. However, CD4 is not on the causal pathway. HIV viral load is on the causal pathway towards benefit in reducing the likelihood of mother-to-child transmission of HIV. An HIV viral load has

been used as a surrogate endpoint in many HIV settings.

Intervention may affect one of multiple pathways, as shown here. And Dr. Walton spoke a little bit about this earlier. If the pathway that it affects is not a primary pathway towards clinical benefit or determining what the clinical outcome is, then having large effects on the biomarker, which would show that your agent is active, may correspond to small effects on clinical outcome.

An intervention may have some other mechanisms of action besides the desired one, intended one, that affect clinical outcome. We often refer to these as off-target effects, and these off-target effects may positively or negatively affect outcome.

For example, erythropoiesis stimulating agents, ESAs, impact hemoglobin in a positive fashion. But they also affect and increase the risk of thrombosis. And that increased risk of thrombosis increases the risk of a cardiovascular

event.

For using a surrogate endpoint to draw conclusions of whether there's maybe a positive or negative effect on a clinical benefit endpoint, from an intervention -- and this corresponds with using a surrogate endpoint for regular approval -- the surrogate endpoint must be correlated with outcome. An entire effect of the surrogate endpoint -- the entire effect on the clinical benefit endpoint must be captured by the effect on the surrogate endpoint. This is known as Prentice's criteria.

One thing that often is misunderstood is when we talk about surrogate endpoints and Prentice's criteria, Prentice's criteria corresponds to drawing conclusions on the clinical benefit endpoint from analysis on the surrogate, much of what we would do in a regular approval.

Assessing from previous studies the reliability of a surrogate endpoint, one way to do so is to investigate the pattern relating effects on the surrogate endpoint with effects on the

clinical benefit endpoint and the amount of deviation from such a pattern.

For example, Dan Sargent did an evaluation of the predictability of comparisons on disease-free survival to comparisons on overall survival in adjuvant colorectal cancer. He considered studies that use five fluorouracil-based chemotherapy. The hazard ratios from those studies for disease-free survival and overall survival are plotted here.

There is, as denoted by this dotted line, a linear behavior between hazard ratios of disease-free survival and the hazard ratios of overall survival. There is some deviation from that pattern, and the typical deviation is a difference in the overall survival hazard ratio of .05. That's how much the overall survival hazard ratio tends to deviate from what you would get on the line, by .05.

Now, it's also the case here, important to note, that large effects on disease-free survival -- these are all experimental over

controlled hazard ratios corresponding to large effects on overall survival. Moderate effects on disease-free survival corresponded to moderate effects on overall survival. And small or negative effects on disease-free survival corresponded to small or negative effects on overall survival.

Formally, that's what makes disease-free survival in the setting of studies evaluated here a good surrogate for overall survival. Results from such a meta-analysis apply to the types of studies you looked at, the therapies and the disease setting. Thank you.

DR. REAMAN: Thanks, Mark.

Dr. Maria Grazia Valsecchi will give a presentation on surrogate endpoints as well.

Presentation – Maria Grazia Valsecchi

DR. VALSECCHI: Yes. Hello. Thanks for inviting me here to talk. I will follow on, on what Mark said. I will try to illustrate some of the problem of understanding the difference between a prognostic biomarker and a surrogate biomarker by using some cartoons.

This is the definition we have seen already twice, and so I will skip it. But only thing I want to underline here -- in red this is written -- that a surrogate marker may be also a treatment-specific surrogate, not necessarily a surrogate for a class of treatments, is a surrogate for all treatments, classes.

In the MRD, the question we have today on the table is this one. So MRD we know is a strong prognostic factor. So it is a prognostic marker as defined today earlier. And of note, the strength of this prognostic marker remains even if MRD is in some trials used to stratify patients; so to tailor treatment intensity based on MRD at a late endpoint. In spite of this, MRD remains a prognostic factor.

The other question we have to ask, following what Mark said, is MRD able to capture fully the treatment effect on the clinical endpoint? And this is not proved yet. And this is what needs to be proved in order to use it as a surrogate.

One question that's behind it -- and Biondi

mentioned it before, and also Steve Hunger -- could it be easier to prove surrogacy for targeted drugs, which have a specific -- let's say a more likely clear mechanism of action than a cytotoxic drug?

As I said, I will use some cartoons to illustrate this. And let's see first what's a perfect surrogate. And I will use the term "activity" to denote the effect of treatment on response.

If I have a standard treatment, an experimental treatment, randomized on 100 patients, then MRD response, for instance here, shows that there is a high activity because the experimental treatment is 60 percent, let's say, of responders versus 40 percent in the standard treatment. So they have a low level of MRD, for instance, or negative MRD.

Now, let's see what happens -- in order for MRD response to be a surrogate for clinical benefit, then we have to look at what happens in terms of failure rate, let's say 40 years failure rate in these groups of responders and non-

responders. And here we have that the failure rate, when we have a perfect surrogate, then conditional on having MRD response. Either in the standard or in the experimental treatment -- it doesn't matter -- we will have a lower failure rate than in the non-responders. And that's the same. That's what we mean when we say that a perfect surrogate captures the effect of treatment. The effect of treatment expresses itself through it. And so we have the same failure rate in non-responders and in responders in the two groups.

Here, we illustrate with this simple cartoon that, at the end, in spite of a high activity, 20 percent difference in activity, we may have a relatively modest effect on clinical outcome, just because of all the numbers combined here.

In this context, in fact, 20 years ago, when the discussion of surrogate endpoints started in solid cancers, especially, that's why people started to realize that they should have then looked at meta-analyses in order to see whether this modest effect was confirmed in many trials and

were worthwhile pursuing.

Now, this is the context of a perfect surrogate, but what if we have not strictly a perfect surrogate -- so something that is somehow -- here, we have again a high activity? The experimental treatment gets more responders than the standard treatment.

Yet, we have here a phenomenon, which may be realistic, that we have a higher activity. But in the 60 percent of patients where we have a high activity, we have also a little bit higher failure rate than in the 40 patients where we have a high activity -- where we have an MRD response in the standard treatment.

This is to say that we can have a situation, the scenario, where we have a high activity but no clinical benefit because, in some ways, the quality, let's say, of this response does not translate into the same effect on the long-term clinical benefit.

So here, we arrive to no clinical benefit, having had high activity. Here, actually, some

patients became responders, but they carry with them a little bit more failure rate. And so we have this situation of not seeing clinical benefit.

This is not a perfect surrogate. And this scenario relates in a way on how you define a cut point that discriminates prognosis. How do you define responders? Are they negative, like at 10 to the minus 6 level, or you adjust with an MRD less than 10 to the minus 4 level? Obviously, in defining the cut point, you will also define the level of activity, you will see, but also the failure rate that these cut points bring in the long term.

The other point we need to discuss is which is the time point where MRD should be measured. And we know that the early time point is of interest because results would be seen in a shortened time, especially in the context of drug development. And we know also that in leukemia, in shorter time points, disease levels are still heterogeneous, while patients tend to become negative the more the time elapses from diagnosis.

Also, we know that, especially in high-risk subgroups, relapses tend to occur earlier, and an early time point may be then quite predictable relapse.

The disadvantages are that, obviously, looking at early time points will limit research treatments of use in early phases. If we want to define surrogates for maintenance treatment, then we have to have later time points. And often, for instance, we have seen an example before in solid tumors. The surrogate becomes, then, the progression-free time, the progression-free survival, for survival as a clinical outcome.

The other disadvantage is that in the majority of patients with ALL, where relapses instead occur late at the end of therapies, early time point may be poorly predictive. And this is an issue more related to the evaluation of treatment strategies rather than drug development, obviously.

The last point I want to make is to show a cartoon of a situation which we indeed encountered

already in the U.K. trial on relapse ALL, where we took mitoxantrone. And also, it was mentioned by Steve Hunger, on the trial related to the comparison between the trial on dexamethasone in front-line ALL because, there, what we saw is that we see that there is no activity, yet there is an effect on clinical outcome.

So how can this happen? Because just MRD response is not a surrogate here, obviously. Yet, it is a strong prognostic factor because the responders have a failure rate which is less than the non-responders, both in the standard and in the experimental treatment. And yet, treatment has an effect on outcome because, obviously, here we see that the experimental treatment has overall less failure rate than the standard treatment.

So this is to illustrate the complicated situation of going from a prognostic factor to a surrogate. And the criteria for validation were already mentioned. We need, obviously, the treatment effects, the surrogate, the treatment effects, the clinical endpoint of interest, being

that survival, or event-free survival, or whatever, that the two are correlated and that the treatment effect in fact is absorbed completely by the surrogate.

DR. REAMAN: Thank you.

Dr. Devidas also will discuss the potential role of MRD as a surrogate.

Presentation – Meenkashi Devidas

DR. DEVIDAS: Good morning. I'm happy to be here and be part of this workshop. Continuing on with the discussion, a lot of what I wanted to say has already been said before, so I'm not going to be repetitive.

To start with, I just want to touch on data from -- the data I'm presenting is all from the Children's Oncology Group. Some of it has already been presented, but I'll just touch on it to clarify a few things.

The initial trial, the 9900 series regarding MRD data at the end of induction on more than 95 percent of patients with pre-B-ALL, there was no intervention other than the standard induction

therapy given prior to measurement of MRD. But as you can see from this outcomes curve, it was highly prognostic, being MRD negative, which was more than 80 percent of patients had a very good outcome.

MRD was not used as part of the risk stratification on that series of trials, but it was highly prognostic, as can be seen here, where we're looking at relapse-free survival. The MRD-negative patients did significantly better and were less likely to relapse than those who were positive.

In a multivariate model, as I think Greg mentioned earlier, the most prognostic factor when you include all the usual clinical factors there, or cytogenetic factors, the most significant factor was MRD end induction. Day 8 MRD was also significant, but not quite as much as day 29 MRD.

On that particular series of trials, we included -- the treatment intensification, in general, was plus or minus, giving a delayed intensification phase. And when that was included in the multivariate model, as you can see up there, day 8 MRD was still the most prognostic factor,

even when treatment was included in there.

I just want to touch a little bit on the recently completed COG trial from the recent series for NCI high-risk patients. Steve Hunger did mention a little bit about it. I just want to touch base on it a little bit more.

It had initially a two-by-two factorial design with two randomizations, one being the steroid question between dexamethasone and prednisone. And the second one was between high-dose methotrexate and Capizzi methotrexate.

Later on during the trial, due to other adverse events, osteonecrosis, et cetera, the steroid randomization was restricted to the younger patients less than 10 years old, while the methotrexate randomization continued in all patient groups.

So basically, the four treatment regimens abbreviated there were PC, PH, DC, and DH, depending on which steroid and which methotrexate arm they were on. Interim analyses in 2011 showed high-dose methotrexate to be much more efficacious

than Capizzi overall, with a five-year EFS of 82 percent compared to 75.

Among the less-than-10-year-olds, there was significant interaction between the two study questions. And so we ended up comparing the four regimens separately. Since high-dose methotrexate was shown to be superior to Capizzi overall, we restricted the comparison to the patients who got high-dose methotrexate when we were comparing the steroids. So DH versus PH was the comparison. And it turns out DH was much more superior to PH for the younger patients, with a five-year EFS of 87 compared to 80 percent.

As Steve pointed out, the MRD rates at the end of induction on that study, if you look at the DH and PH arms, were 81 versus 83 percent, pretty much the same. And if we had actually thought of using MRD as a surrogate in that study, we wouldn't have caught on to the fact that high-dose methotrexate was actually efficacious and gotten those results then.

So clearly, it's not a surrogate. On this

1 particular study, actually, MRD was used as part of
2 the risk stratification. It was some point. We
3 got rid of patients who had very high MRD levels
4 very early on. It's clearly not a surrogate, but
5 it is highly prognostic, both with overall and also
6 within the treatment regimens.
7 Here is a graph showing the -- actually, if
8 you look at the MRD-negative patients, they
9 have -- I don't have the actual EFS rates on there,
10 but they were significantly better than the rest.
11 So then the question is, MRD, do we not
12 think of it as a surrogate or should we?
13 Obviously, in the general ALL population, it may
14 not be a good idea to use it as a surrogate. We
15 don't know. But as Maria Grazia pointed out, at
16 least in pediatrics, the way we design the trials
17 for frontline ALL induction, we haven't messed
18 around too much with it. And it's in the first
19 month of therapy. And at the end of the first
20 month is when we are measuring MRD.
21 So it's a little difficult to use that early
22 time point, use it as a surrogate for efficacy at

the end of a trial where therapy can go on for more than two years, in a population where most of the events tend to occur later on. And most of our interventions tend to be post-induction. So it's a little difficult to do that there.

But there might be other subsets that we could use MRD as a surrogate. Questions to be thought of is, what is the appropriate time point? This has all been touched on before. The sensitivity levels; should we go with .01 percent, as has been specified here, or should we go with the higher cutoff?

What kind of subsets can we look at? Maybe higher-risk ALL subsets like relapsed ALL, infants, maybe other patients with various high-risk markers, like Ph-positive ALL, JAK mutations. There are a whole bunch of new markers coming out now.

In most of these subpopulations, people are looking at targeted therapies. And the therapies may be given early on. The interventions are early on. And we can have appropriate early time points

for MRD, which might be suitable to assess efficacy or activity. So these might be appropriate subgroups where we might be able to actually show that MRD is a surrogate for long-term outcomes.

So if we look at early-phase trials for these targeted therapies, we can introduce new agents early on in therapy, induction, or whatever else you want to call them, phases, assess activity or efficacy using MRD as a surrogate for EFS, or overall survival, or progression-free survival as the case may be.

We can also use it as a method as an early endpoint in screening trials, which are becoming more popular, where we are screening for, say, looking at more than one agent in a certain population. And we are able to get results faster and maybe move the drug forward a little faster than having to wait for long-term outcomes.

So as an example, it turned out, as a coincidence, that the most recent -- the Journal of Biometrics, the most recent issue had a whole series of discussion articles on surrogate

endpoints. This was in March, so I didn't quite have time to go through all of them. But as an example -- I just wanted to talk about one of them.

So basically, if we're going to be using MRD as a surrogate in these subpopulations, I guess we do have to come up with maybe some new novel ways of designing some of these trials. And that's pretty much what they were focusing on, maybe come up with some adaptive trial designs using surrogates.

The example that I am focusing on is the paper by Renfro, Carlin, and Sargent on Bayesian adaptive trial designs for a newly-validated surrogate endpoint. I'm not going into the gory details of it, just touching on what they are talking about in it.

It allows for a new surrogate endpoint to be primary in assessing the effect of an intervention. This particular endpoint was validated as a surrogate, based on multi-trial historical data, which showed a relationship between the surrogate and the clinical endpoint.

So what they suggest is you evaluate accumulating data as the trial progresses against that relationship, assuming it's a surrogate. And this way, you are guarding against any erroneous assessment of treatment effect, based on a truly invalid surrogate.

When the joint outcomes on the new trial seem to be now in line with what was seen on the historical trials, we could proceed with the surrogate endpoint as the primary endpoint and, adaptively, they do include rules then for stopping the study early, for early evidence of success, inferiority of an experimental regimen, futility rules, et cetera. If there's early evidence that the surrogate needs to be discarded because it doesn't quite fit in with what was seen on historical trials, then you discard the surrogate and just use the regular primary endpoint, use the adaptive rules to the original primary endpoint.

With that, I will stop. Thank you.

Clarifying Questions

DR. REAMAN: Thank you.

We might be able to open this session up for a few questions for our statistical presenters, if there are questions from the panel at all.

(No response.)

DR. REAMAN: I guess I just would ask the question, back to Marc or Liz, within the perspective -- or from the perspective of context of use for a surrogate.

Is that something that has to be fixed, or can that context of use change within a disease that it's being used in as a surrogate endpoint, based on situations; if we're talking about looking at MRD as a surrogate in high-risk patients, particularly in the relapse setting, where a specific intervention is the addition of a new drug versus in the newly-diagnosed setting, where there is really sort of comparative evaluation of the efficacy of regimens?

DR. WALTON: Within the framework of formal qualification, we talk about the context of use being a very precise description of how one uses the biomarker and what conclusions one draws from

it.

So we very much expect that a particular, specific situation may call for the biomarker to be used in one specific particular way to allow a certain kind of conclusion.

A different situation, a different circumstance, a different set of patients, a different stage of the disease, prior history of treatments, any of those things, may change the way in which we need to use the biomarker and to apply it.

So I think it's very plausible that we may have different ways of using the biomarker in different situations. We just need to have enough information to justify each of those particular ways of using it in each of those particular situations.

DR. HUNGER: I had a question for Dr. Rothmann. The slide you showed on the correlation between disease-free and overall survival hazard reductions in various trials of, as I understood it, colorectal cancer treated with

5 FU-based regimens, I thought the data were striking. And I wonder how the agency interprets those.

Would you interpret them that disease-free survival is a valid surrogate in colorectal cancer, based on fluorouracil-based regimens in colorectal cancer in general? In any cancer?

DR. ROTHMANN: I guess, first, technically, the results are just good for the types of trials that were looked at, which had an experimental arm, surgery plus a 5 FU with maybe methotrexate, or vamasol, or leucovorin.

Dan Sargent did present his results at an ODAC. I know sometime thereafter, we did approve oxaliplatin for adjuvant colorectal cancer based on a comparison on disease-free survival.

DR. PAZDUR: In general, we have used disease-free survival as an endpoint for a regulatory approval, not only in colon cancer, but in other solid tumors such as breast cancer.

DR. HUNGER: Thank you.

DR. REAMAN: Dr. Borowitz?

DR. BOROWITZ: I'd like to ask Dr. Valsecchi a question. I was intrigued by your slide where you constructed the perfect surrogate, in which you had a significant effect on the surrogate, MRD in this case, and only a modest effect on outcome. And that was an aspect of this that I've actually not thought of to date.

I want to get a sense of how common -- with reasonable assumptions about activity of a drug, with both a surrogate and in a patient population, you know, reasonable assumptions about recurrence rates and so forth.

Was this a very contrived situation, that you had to really plug in certain kinds of numbers? Or is that actually going to be something that we're going to be up against a lot, even if we find a perfect surrogate?

DR. VALSECCHI: This is constructed to show the problem because this problem was raised in solid cancers, like in breast cancer, years ago, when response in terms of tumor shrinkage was not really related to an evident clinical benefit.

So this has been shown to illustrate that this can also be the case. Obviously, the higher is the activity, so the higher the difference in activity, even if the play of numbers in terms of failure rate is similar to what I showed, the distance in the clinical outcome becomes higher.

But this was a problem which I believe -- I don't work on breast cancer, but I believe it's still a question mark in many issues. So that's why I showed the context.

DR. REAMAN: I think maybe we'll go onto the next session and discuss some of the technical issues and considerations. I'm going to ask Dr. Bagg to introduce the speakers. We may have to change the order of our speakers because of some travel difficulties, but we'll see what happens.

Dr. Bagg?

DR. BAGG: I also just wanted to use this opportunity to remind, if I may, some of the European panelists and some of the other people here who are not associated with the FDA, if you'll recall from a couple of the first talks, a lot of

different abbreviations were used that you may not be familiar with, AIEOP, LDT. I hope you were paying attention because there will be a test at the very end.

(Laughter.)

DR. BAGG: The first speaker is Dr. Jacques van Dongen from Rotterdam, who is going to give us an overview of his enormous expertise in the area. His title is very long. I'm not going to repeat it.

Presentation – Jacques van Dongen

DR. VAN DONGEN: Thank you very much. It's a pleasure to come over -- to be invited by Gregory Reaman and Steve Hunger -- and share our experiences from the other side of the ocean on MRD diagnostics.

So when we talk about MRD -- and this is the measurement of low numbers of leukemic cells -- then we mean that we try to measure very low levels. And in the meantime, it's clear from many studies, not only the European ones, also the ones in the U.S., that we definitely need to reach

10 to the minus 4. 10 to the minus 4, what we say now in a quantitative way, and even pitch 10 to the minus 5 to be more accurate. And the technologies, so this is what you would like to measure, that is leukemic cells.

The methods that have been proposed over the last 20 years are summarized here. So that's flow cytometry. That is clear that you measure cells. That's completely different from the PCR analysis for immunoglobulin and T-cell receptor genes, where we measure DNA, one copy per cell only.

Finally, last, about fusion transcripts, which is used a lot in chronic myeloid leukemia for monitoring Bcr-Abl fusion genes. And then we measure multiple copies per cell, again, completely different things that we are measuring. And we can of course assume that you get different results, and that's the case.

Sorry about the size. My computer is different than the computer used here.

So what was decided in Europe? Well, at the time that the large study started, it was this

focal analysis, and we saw that we couldn't easily reach 10 to the minus 4. So in Europe, it was decided to go for the PCR analysis by immunoglobulin genes to reach at least 10 to the minus 4, preferably more sensitive.

A lot of discussion was there, so a lot about definitions. What means MRD remission? How should it be measured? What is the sensitivity to be reached? When should we measure? After induction treatment? Should it be continuous monitoring? Which we did, in fact. In the start of all the studies, treatment blocks. And the time points, but also the quality of range. Would that be different per time point? And again, the discussion about flow cytometry and PCR is even still ongoing.

What was very clear is that standardization is really needed and that could only be achieved by networking. And these are the topics that I would like to address in my presentation.

So the first results that we go to are based on immunoglobulin genes, and we have to realize

that in these immunoglobulin genes, these junctional regions are the targets. So we are talking about leukemic cells, extract the DNA, and find these specific targets, a diagnosis.

That has to be done for each individual patient, meaning that's a lot of work, and there is potential also in the biological variance, because not each rearrangement is the same, so your sense might be slightly different. And you could also have leukemic clones that are different in different patients.

Nevertheless, based on doing this -- and this was started in the '90s -- this was the result that we got, end of the '90s, where we had the low-risk group identified, based, by the way, on two time points. The first time point is after one month, the second time point after three months. So the whole induction treatment is included in the evaluation of MRD.

Then there is this high-risk group and this intermediate-risk group. The high-risk group has, indeed, very high levels, at least 10 to the

minus 3. And you see this enormous gradient, which was striking for us at that time, because this was a complete novel concept. But the most striking was that if you look to where the relapses are, 45 percent of all the relapses in the protocol were in the high risk. Most of those were in the intermediate risk, and only very relapses, very few of the total relapses were in the low-risk group.

So that would open possibilities not only for treatment increase and treatment adaptation, but also reduction of treatment, and that was an important point.

Nevertheless, we realized even more, when we started to use the real-time qualitative PCR assay, so mainly the Check Point technology, that we needed high levels of standardization. And just to give you an example, if you would provide exactly the same electronic file, and ask the different laboratories to please analyze our data, and see what comes out, they only got 70 percent concordance because we didn't talk about cutoff. We didn't talk about how to interpret the data.

Then we started to make guidelines. We made guidelines and we talked about what is qualitative range, what is background. So we defined everything. This took us more than four years, close to five years, and 10 quality control rounds to agree on how to interpret the data.

The moment we have done so, in a new MRD, where all this work has been taking place, these are the laboratories that collaborate. There are now 43 laboratories in 18 countries collaborating on this already for 10 years. We have, last November, the 20th quality control round. And now, indeed, finally, we reach a concordance based on the same files of more than 95 percent, close to 97 percent, exactly the same interpretation of the data. I'm not still discussing about how to create the data.

So the point is standardization and making agreement, making very strict guidelines, very strict, is critically important. And that brought the PCR technology really forward. And I can say that Europe and the countries that are

participating -- that is around 400 million people -- all the children in those countries are treated according to exactly the same MRD-based strategy, which is a major step forward compared to 10 years ago.

Did it work? Well, this is one example. This was published two years ago. That is about the most complicated group on ALL, I would say. That's the infant ALL, very poor prognosis. And here, you see that this standardization worked. This one is also worked from different laboratories. Again, an enormous gradience between risk groups and, again, the recognition of a low-risk group, now with more relapses.

Thirteen percent of all the relapses are occurring in this low-risk group. Still, it's clear that it can be done. You can collaborate and create data that are really comparable between laboratories so that also internationally you can exchange the results.

However, it's a lot of work. So using this technology, the immunoglobulin gene-based -- T cell

receptor gene-based approach is really labor intensive. It's about identifying, per patient, a diagnosis, the rearrangements you require, a lot of knowledge, specialized laboratories. It's time consuming. Already, the target identification and the assessment of your sensitivity takes four to six weeks. That is an enormous amount of work. And logically, the question is then, can this not be done by other technologies, for instance, flow cytometry?

Now, many labs have tried flow cytometry. And the largest study that I am aware of is from the United States. This is the third time that you've seen this graph. This from the COG. And Mike Borowitz has published this. And it looks like an enormous gradient as well.

There's one big difference with the PCR-based studies, and that is very large low-risk group. There's a very large low-risk group. It's 80 percent of all the patients and contains more than half of all the relapses in the total protocol.

So it depends on which technology you use. This technology, flow cytometry, as presented in this slide, is perfect for recognizing the different types of high-risk groups. It is not really good for recognizing the low-risk patients. So it's really depending on what is your aim for MRD diagnostics and what do you like to reach.

If you only like to recognize your high-risk patients, flow cytometry will work fine. If you would like to recognize all the patients accurately, also the low-risk patients, and reduce treatment, as we do in the Netherlands now, then it won't work.

So what we are talking about is what is your MRD window? Your technique determines this and is depending on where you measure, because in the previous slides, this is the measurements on day 29. What I showed to you, and for the European studies, it's measuring at one month and at three months, so much more treatment is evaluated.

So measuring at one week, two weeks, or months, or even at three months, that's really

determining a lot. And then of course, is your technology really going down below 10 to the minus 4, really below that? That's the minimum to reach, in a quantitative way, so it's not pitching 10 to the minus 4; it's going down. Or are you going to use the flow cytometry, where the difficulties and the pitfalls are different. That's what we have to realize.

The enormous advantage of flow cytometry is, they are very high speed, and you don't need to do an enormous analysis and diagnosis. So for all kinds of international studies, you would prefer flow cytometry. The problem is that, in the flow cytometry, particularly for precursor B cells, there's this enormous background. Now, you are looking to a monitoring process during and after treatment, not to looking at the malignant cells, but the normal precursor B cells.

The pink ones are the early normal precursor B cells, and the yellow-green ones are the late precursor B cells. This enormous outgrowth, regeneration of precursor B cells, that's one of

the most complicated factors. And that's exactly at the time of one month and three months after start of treatment. That's all of the main complicated factors.

It means that the normal precursor B cells, the regenerating normal precursor B cells, give a problem, but not only that, also the leukemic cells, not in each patient, but in many patients, particularly in the corticosteroids, show a particular phenomenon.

Here you see the diagnostics at day 15 and day 28, leukemic cell population in red. And what do we see? If you look to CD10 on the Y axis, at CD20, here below on the Y axis, then you see that it's going to decrease. It means that there is an immunophenotypic shift of both things. The normal regenerating cells and some changes in immunophenotype might play a role in the sensitivity of the assay.

Then, of course, everybody is comparing results. And there are around 10 -- I can show you 10 of these graphs in the literature about how well

PCR or flow cytometry can be compared. Here it is on the horizontal axis and RQ PCR on the vertical axis. And what do we see here? There's a lot of samples, positive and real-time quantitative PCR that are negative by flow cytometry.

This is one example and a quite recent example. This is from the Irish group. Now, it's the Scandinavian group, did a comparable study. And now RQ PCR is on the horizontal axis and the flow cytometry on the vertical axis.

What do we see? The same problem. There is many samples that are clearly positive, and I would say nicely quantifiable. This is all based on the same Euro MRD criteria, the PCR, because the Scandinavian groups are included as well, and Euro MRD, so they have the same guidelines. So this is really in the quantitative range. And what do we see here? That's much less here.

These red points are the day 29 samples. So that's quite critical. And there is more than one log difference for many of these red points.

This is a study that we just finished. Now,

we only provided a comparison between RQ PCR and flow cytometry in the qualitative range. So we left out everything that was, let's say, not quantifiable or negative, just to show again how it looks. It looks fine, but again, you see, based on this cloud of points, that RQ PCR is apparently more sensitive. Now, you might argue, is this going to influence our risk groups?

If you use exactly the same criteria for PCR and flow cytometry on the same type of patients -- then I show you these results, this is a 4-color approach, a 6-color approach, and real-time quantitative PCR at day 15, day 33, and at three months. And, oh, that looks very comparable. You say, oh, that's very comparable. And now you look to these data. Well, it's not too bad, you would think. And at day 87 (sic), it looks also not too bad. Well, you see clearly the decrease for 15, 33 to day 78. The green, of course, are the negatives. The reds are the highly positive ones.

But this is the problem. PCR more carefully

identifies these groups, these low positive groups. That means they are positive at a low level or at a low level around 10 to the minus 4 or slightly lower. These are the groups that are different.

Why would this be different overall? If you now look to the same set of data, this is from the ongoing ALL 10 protocol, where we have no more than 550 patients included, and it was randomized with both methods.

Now, if you look at both methods and look over the risk group definition as it is used in the protocol now, MRD-based, we have very limited high risk. Then this is the medium-risk group and the relatively small low-risk group.

By the way, the low-risk group only has 2 percent relapses. And the treatment reduction referred to all the anthracyclines were removed from block 2. That's a very heavy cytotoxic treatment reduction with the small group. That's clear, this low-risk group. And then look to the flow cytometry-based approach. There you see that's completely different, and the main problem

is here.

So there are many PCR-based medium-risk patients that are clustered in the low-risk group by flow cytometry. But these patients obviously have not been regarded as low risk, but as medium risk in this treatment protocol because the PCR base was the guideline for the stratification.

So the point I would like to make is, it's about what are you evaluating? Are you only evaluating a diagnosis, or maybe at day 15 and day 33, day 29 by flow cytometry? Then you definitely analyze a part of the corticosteroids or the full induction is analyzed. But this is -- we end week 12, so that's one and three months.

It's really depending on where you measure, and what you measure, and how sensitive you measure. That determines the different composition of the risk groups. And if you look particularly to not the high-risk group -- the high-risk group is very comparable. It is very comparable between the different technologies.

If you go to the medium-risk group and the

low-risk group, there is a major shift between the groups of 25 to 40 percent, depending on the treatment protocol and maybe also depending on the type of technology.

So if I would like to choose in the near future, I would like to go for flow cytometry because of its speed, because I believe it can be international and more easily standardized, but still, a lot has to be done because is indeed flow measuring the same as PCR? I don't think so. I mean, definitely, one is DNA, a single copy per cell, and the other is cell population, so you measure something else.

You can easily replace it. I would say not now. You could supplement potentially because you saw in the previous slide that there were some cases that we did not analyze, and clustered by molecular technologies. So there you could decide to use flow cytometry, so supplementing maybe, just depending on what you decide for your protocol.

The advantage of flow cytometry is it's very broadly available. You don't need special

facilities. There is very good equipment available for more than 8-color flow cytometry at I would say a still moderate cost.

But standardization, improvement of particularly the sensitivity is critical in my opinion. It has to be broadly accepted in terms of the antibody protocols. You cannot just say, "I use a particular antibody." It has to be precisely defined because, otherwise, your results will be different. And you need to have international guidelines for data acquisition. But not only that, as I explained for PCR, also for the interpretation of the data, that has to be fully standardized.

So if you would go for that, really, for a multicenter design, then you need the standardization. And it means that you have to innovate your flow cytometry a lot. I believe we have to switch to 8 colors or even more to increase your sensitivity. Potentially, in addition, for new markers help.

Software is critically important because

handling an 8-color approach needs a completely different type of software that is not generally available by the large companies. So that's why we handled it in EuroFlow and designed a completely new software for fast and automated analysis.

It means that you have to evaluate everything in parallel. And the data analysis and the guidelines for standardization should also include instrument setting, data acquirements, and data interpretation.

So just to give you an example of where it is possible, this is flow cytometry. Now, you are not looking to a normal dot plot. This is a principal component analysis. This is an end-dimensional -- in fact this is a 10-dimensional view on flow cytometry, where the leukemic cells of course are in red. And these are different times of regenerating precursor B cells from different types of bone marrows in different phases of treatment.

So different phases of treatment give you a different type of regeneration in your precursor B

cell compartment. And you can separate, indeed, your leukemic cells of this single case.

Does it work further? Well, we decided -- because this approach, as I show you now, was not feasible in all the patients. Then we decided to go for a much better insight, and what is the normal maturation pathway of precursor B cells as the counterpart of precursor B-ALL.

Again, you're looking to what is called an APS view, not a normal dot plot. This is principal component analysis, where we recognize 20 differentiation stages. And now, for us, it was critically important to understand whether regenerating bone marrow is different or comparable to normal bone marrow.

Here, you see the results. In green is the regenerating bone marrow of this complicated stage at day 29 and at also day 30. At three months, you have comparable results. So you could say logically -- as an immunologist, I would say this is logical, that normal precursor B cells that regenerate fall together as the normal pathway.

This was never proven before because you could think also that there are some aberrations.

If you now plot leukemic cells against this pathway, then you see you can plot here in red the leukemic cells. In most of the cases, they're different. And I could use different views.

For instance, these are the four cases. This falls together. If you go to a second APS view, now they fall aside. So depending on the type of leukemia, you can still look at the leukemic cells.

It's not yet solved because you still encounter some problems, but there is an enormous increase in knowledge now, based on 8-color flow cytometry, not analyzing in classical dot plots, but in an array that you, in effect, use principal component analysis, all the data in one step.

This is fully standardized. These data are derived from 10 different laboratories and you can overlay the data completely. So we think that flow cytometry, potentially even more than PCR, can be standardized. It needs, however, an enormous

effort.

So what I would like to conclude is, so far, the gold standard in the European protocols has been the PCR technology for IG/TCR genes. I guess this will stay so for the next maybe few years. Of course, there is an enormous difference in MRD value between the protocols. It's also depending on -- and my guess is that the difference that we see in the literature is mainly based on the fact that the MRD techniques are not standardized and the fact that they differ in sensitivity. These are the main reasons why we see differences in MRD value between protocols.

As I already said, I hope and I guess it will be possible to replace the PCR-based technology by 8-color flow cytometry. This requires novel developments, but particularly it requires standardization. It needs the quality control rounds and a lot of international collaboration to get this done.

So that's what I believe that this green conclusion is what we have to go for. Thank you

very much.

DR. BAGG: Our next speaker is Brent Wood, who is physically here and arrived a few moments ago from the University of Washington.

Are you ready to go? Okay. You're up.

Presentation – Brent Wood

DR. WOOD: First, I'd like to thank Delta Airlines for my timely arrival to this meeting.

(Laughter.)

DR. WOOD: What I wanted to focus on are mainly kind of technical issues related to flow cytometric assessment of MRD. Many of these have been covered by Jacques in the previous talk, and so we can move through these I think fairly quickly. But the idea is to give you some idea as to how this technology is employed in the assessment of minimal residual disease and how well or less well standardized it is at the current time.

So the first basic problem in performing an MRD assessment using this technology is the actual ability to identify the presence of populations of

cells and distinguish them from their normal counterparts. And this relies to a large degree on the fact that normal expression of proteins or antigens by a cell is driven by an underlying genetic program, which drives cell differentiation and maturation. This results in a consistent pattern of protein expression on the surface of the cell, as well as inside the cell that's essentially invariant between individuals, regardless of their age, sex, ethnicity, et cetera, so that antigens are expressed in a consistent and reproducible pattern with maturation.

When these cells become neoplastic, the genetic program is altered in a variety of ways that we're now just beginning I think to uncover, and this results in changes in the expression of proteins of a variety of different sorts. The basic categories are listed here. But the idea is that it allows us to distinguish these abnormal populations of cells from normal populations because of changes in the expression of proteins, as a result of underlying genetic abnormality.

One can then go through, and, for any given antigen or protein that's being expressed, catalog its expression with regards to maturational stage, and create diagrams of this sort -- and there are many of these that have been generated over the years -- that describe the expression of a given antigen on a given lineage of cells, in this case B cells, during specific maturational stages.

One can see that each of these different antigens is expressed in a characteristic pattern within normal B cell development. One can use this information, then, to construct panels of reagents that would allow us to identify the specific stages of maturation. And as it turns out, many of these same antigens are also aberrantly expressed when these cells undergo neoplastic transformation.

So for instance, this is an older kind of 8-color flow cytometric dot plot, showing a variety of different projections of data, focused on the B cell population gated on CD19 expression and showing how there is a very tight correlation between the expression of many of these different

antigens throughout the maturational sequence, starting from the most immature populations, colored in yellow, to the next stage of maturation in light blue or kind of aqua, and toward more mature forms, that are colored either blue or red.

So in normal individuals, this pattern of expression that's being depicted here is essentially invariant.

If one then looks at patient samples that have been treated for ALL, one can then identify the residual population of leukemic blasts by its difference from this normal immunophenotypic profile that we've come to understand quite well.

In this particular example, this is a medical student from our institution who developed ALL, was treated, and now is deceased, and at the time of end-induction therapy, had an abnormal immature B cell population of about 0.1 percent, which is pictured in red.

It is very distinct in terms of immunophenotype from the other populations of cells, either immature B cells in light blue or

more mature B cells in dark blue. In particular, there are differences in expression of CD10, which is over-expressed. Its relationship between CD10 and CD20 expression is aberrant. CD45 has been essentially lost on this population in comparison to normal. CD34 has been down-regulated. CD58 is over-expressed.

So in many cases, it's relatively straightforward, using an appropriately informative combination of reagents to identify these abnormal populations of cells and distinguish them from either normal mature or immature B cell progenitors that are present in bone marrow specimens.

This whole technology relies on the idea that the immunophenotype itself is actually sufficiently informative to provide this kind of discrimination. And this is where, as Jacques mentioned, there is some work or opportunity to be done to identify reagents that might actually provide an even greater degree of separation of normal from abnormal.

Here, we have two examples of ALL, just as a

way to demonstrate how more or less informative a given antigen combination might be. On the top row is one patient where the population of leukemic cells is in red and normal immature B cell precursors are in cyan or light blue. And one can see that the down regulation of CD45, with some up regulation of CD19, allows pretty good discrimination of this population from the normal B cell precursors. CD38 expression, while slightly lower, and CD10 expression, while slightly higher, are less informative in making that kind of distinction.

In the second patient on the bottom row, the informative content of CD45 and 19 is very similar, but because of a more marked down regulation of CD38 expression and a greater up regulation of CD10 expression, the population is now completely separate. And this combination for this particular patient is highly informative and allows very ready discrimination of leukemic cells from normal immature B cell precursors.

As Jacques alluded to, depending on the

stage of therapy that one investigates MRD, the normal number of these B cell precursors may be quite variable. And this directly affects the sensitivity of this assay.

So by taking a small number of these normal B cell precursors and mixing them with a small number of the leukemic population electronically on the computer, one can show this effect.

So in the third column from the left, one can see that in the presence of a small number of immature B cell precursors, in both patients, the combination of antigens is sufficiently informative that one can discriminate with a high degree of certainty the leukemic population from the normal immature B cell precursors.

However, as the number of these immature B cell precursors increases, for instance, what happens later in therapy, the outliers of this distribution begin to obscure the population of interest, and this combination is insufficiently informative for the first patient, but remains highly informative for the second patient.

So for the second patient, the sensitivity of the assay is now largely limited by how many cells one puts through the instrument, whereas in the first case, it's been limited already by the presence of large numbers of hematogones, with an insufficiently informative immunophenotype. So having the informative immunophenotype is key for this type of analysis to be successful.

As a result, people have looked at many different combinations of reagents and tabulated results, published these over the years, describing which immunophenotypes appear to be more informative than others for specific subtypes of ALL, either adult or pediatric.

This is one example that's pictured here. One can see that there's a wide range. Some immunophenotypes are essentially uninformative, while other immunophenotypes, such as CD38 expression at the top, are highly informative.

So having an informative antibody combination is really crucial for this type of analysis.

A number of years ago, Mike Borowitz, as part of the Children's Oncology Group, developed a relatively simple combination of reagents that appeared to be relatively effective in identifying MRD from patients with B-ALL.

This was a two-tube panel, as is pictured here. And both tube 1 and tube 2 of this panel showed a fairly high ability to discriminate leukemic cells from normal B cell precursors. And these are the two tubes together, captured essentially almost all cases that were detected in this particular study.

So one doesn't need necessarily a very highly sophisticated combination of reagents in order to do MRD detection, but it depends a lot on the disease that you're talking about and the time point at which things are being measured. In this case, the measurement was being made at the end of induction, where the number of normal B cell precursors is very low, and so it's a relatively simpler problem.

More recently, this type of approach has

been modified, and this represents the current panel of reagents that are being used in the ongoing COG trials, which is a 6-color flow cytometric assay, collecting about three-quarters of a million events. And in our experience, this type of panel is adequate for the detection of MRD down to a level of .01 percent or below in the vast majority of patients, approaching 99 plus percent of patients at the end of induction therapy, where in these protocols, there are essentially almost no normal B cell precursors left.

So one, again, doesn't need necessarily a very sophisticated approach if one looks at a particular time point.

If one moves to other types of diseases, such as acute myeloid leukemia, the problem is more difficult. And this is in part because the immunophenotypes that are present are much more varied due to the normal maturation, which is much more complex, that occurs from stem cells as they move to lineage-committed progenitors.

So a very common strategy for identifying

populations of cells involves using CD45 and side-scatter gating to identify progenitor populations in the areas indicated here, and then looking at those areas for expressions of antigens in a very similar fashion to what we've just described in B-ALL.

But as you can see, the problem is much more complex. This is gated on that blaster progenitor area, now using 10-color flow cytometry, looking at the expression of many different antigens that are expressed during early maturational stages of progenitors.

The different lineages are colored different colors here. And you can see that there is a lot of complexity to this early maturational stage in the bone marrow, which is where AML is, in part, derived from. So it's a more difficult problem in part because of the increased complexity that's present.

But one can go through and map out, for instance, granulocytic maturation, monocytic maturation, erythroid B-cell differentiation, et

cetera, and use this as the normal map from which we then look for difference or deviation from normal as a marker of neoplastic disease.

So people have looked at AML in a similar way that's been done for B-ALL and described combinations of antigens that are informative for MRD detection. In this case, this is a very large dataset from the Germany group, from Dr. Kern in Munich, showing that, in general in AML, there also are abnormalities in immunophenotype, usually more than one, that allow for the identification of leukemic blasts with some level of sensitivity.

Now, the way in which this immunophenotype data is commonly applied is usually used in two basic ways, one of which has been described as leukemia-associated immunophenotype, or LAIP type of approach. And the way this approach has often been used is that it diagnoses. One runs a combination of reagents, identifies immunophenotype deviation relative to your knowledge of the normal patterns of expression of these antigens.

The subset of reagents that were informative

is then selected or perhaps even a custom set is created of those reagents for use, then, in post-therapeutic monitoring.

Also at diagnosis, a gate is created in those areas that are thought to allow discrimination of the leukemic cells from the normal cells. And at follow-up, this reduced reagent set is run, and events that are present within this area, that's been defined pre-therapy, are counted, and considered to be MRD.

Another approach to this problem is to use deviation from normal maturation, kind of as I've described, in that at diagnosis, one runs the combination of reagents, which is usually a relatively standardized or uniform panel. Immunophenotype deviation relative to normal counterparts is identified, but the exact same combination of reagents is then run post-therapy.

One attempts to identify discreet populations of cells or groups of events that are present within the data, that are different from normal, not using a predefined gate prior to

therapy, but looking at a difference from what one knows to be the normal distribution of these antigens in either normal or regenerating bone marrow.

The diagnostic immunophenotype can be used as a starting point for these analyses, but it's not relied upon as it is in the traditional LAIP kind of approach.

Now, the LAIP approach can work in certain circumstances, as has been published by a number of European groups. This is one example. In the middle row, one can see an abnormal population of myeloid blasts that co-express CD117 and CD2 aberrantly.

As one goes through therapy, a subset of cells remain in this predefined gate, and eventually the patient relapses, whereas in the bottom row, there's a population of blasts that abnormally co-express CD117 and CD7. The events in this predefined gate are gradually reduced by therapy so that, toward after the third cycle, there's essentially nothing left, and the patient

is in remission.

So this approach can work. However, it can also look like this, where the LAIP is defined in red. And one can see, on the right-hand side of the diagram, red events, which appear to be just the tails of normal distributions of cells, because they were felt to be aberrant at the time of diagnosis in this defined area, they're considered as MRD when, in fact, they probably represent just normal counterparts that happen to be trailing into the gate or box that was created pre-therapy, because the focus here is on immunophenotype, not on identification of populations that differ from normal.

Related to this problem is a problem that Jacques mentioned, which is immunophenotypic stability. And it's well recognized now that, in leukemias of many types, these immunophenotypes are in fact not stable.

Here are two examples of ALL. The top row and the bottom row are two different patients. The pre-treatment leukemic sample is in red. The MRD

sample at day 29 is in blue, dark blue. And one can see that there's been a change in immunophenotype in both cases, with a decrease in the level of CD10, and acquisition of CD20 in the case of the leukemia on the top row, an acquisition of CD45 et cetera, a loss of CD34 almost entirely.

So these immunophenotypes are not stable, particularly at these early time points after therapy, so that if one was to draw a box around the areas that are thought to be abnormal in an LAIP kind of approach, one can see that the MRD population largely lies outside of that box and a false-negative result is generated.

We've identified the same thing that happened in T-ALL. This is in a publication from the COG group, showing that immature antigens like CD99 or TDT are commonly down-regulated at these early time points after therapy, particularly by day 29, and, as such, are not reliable markers for the detection of MRD, despite what has been published by some in the literature.

If one looks at other novel antigens, which

we've been screening many hundreds of different antibodies in an effort to identify such antigens, for T-ALL, CD48 is one that we recently identified. And one can see that it quite readily discriminates normal, mature T cells, which are the light blue dots with a higher level of intensity, from almost all cases of T-ALL, which are the somewhat darker blue dots, with a lower level of intensity, with the exception of a very small number of cases that are still different, but closer to normal.

So at diagnosis, it looks like a fantastic reagent for identifying MRD. If one then looks at post-therapeutic specimens at day 29 after therapy, what one finds is that the intensity of this antigen doesn't remain entirely constant. In some cases, it goes up. In some cases, it goes down. In many cases, it remains the same. It's still a useful agent for MRD, but it's not as reliable as would have been predicted by simply looking at pre-treatment immunophenotypes, which is kind of what the LAIP approach relies upon.

The same is true of AML. There are a number

of publications -- found this in the literature; this is just one example -- showing the abnormal blast population in red. And diagnosis has a particular immunophenotype, which as one goes to day 16 after therapy, is already exhibiting marked decreases in the expression of CD117, acquisition of CD15, losses of CD13 and CD33.

This persists through day 32, even though the blast population continues to decline in number, but the immunophenotype remains quite distinctly different from what it was at diagnosis in many respects.

But if one looks at these two different approaches do doing flow cytometric MRD detection, the leukemia-associated immunophenotypic approach has some advantages in that it's conceptually simple and appears on the surface to be relatively objective.

You can define a box in a very consistent manner, and it doesn't require a lot of thought in the post-therapy samples to identify the MRD. Also, if one reduces the reagents that are used in

that panel, we might be able to save a little bit of money.

The disadvantages are that it requires a pre-treatment sample to define the LAIP. If you don't have a pre-treatment sample, you can't do this analysis at all. It also requires immunophenotype stability sufficient so that it can be identified as MRD. And in the event that it shows up in these predefined gates, it is judged to be MRD, regardless of whether it really looks like a discreet population or not.

So there's a risk of false-positive results as well.

The other approach using deviation from patterns of normal maturation has some advantages in that it does not require the pre-treatment sample, because our reference point is normal maturation. A uniform reagent combination is generally utilized for both pre-treatment and post-treatment analysis, so there's some standardization or uniformity that can be achieved.

There's improved specificity through the

identification of discreet populations of cells, not just things that happen to lie in a particular area of multidimensional space. And it's also less sensitive to immunophenotypic instability because, even though the phenotypes change, they rarely normalize and look exactly like normal precursors.

There are some disadvantages, however. And one of the major ones is that it requires a detailed knowledge of immunophenotyping. It takes an expert to look at the data and to be able to identify these populations. And so it's a more difficult thing to generalize to laboratories that have less experience.

There is some subjectivity involved in terms of identifying the populations, because you're looking for a clustering of events and dot plots. And it tends to be, as a result, more time consuming because someone has to sit there and look at these dot plots, and make these types of determinations.

But nevertheless, we think that this approach of deviation from normal maturation has

sufficient advantages that relying on the leukemia-associated immunophenotypic approach alone is probably not a wise idea.

In actual practice, I think many laboratories do some combination of the two in varying proportions, mixing the ability to identify these abnormal phenotypes with their knowledge of normal maturation. And the two together can be, I think, quite powerful.

As Jacques has already mentioned, we've kind of illustrated the time point at which one looks at MRD can be very important. And I won't reiterate this any further.

So this allows us, then, to identify populations of cells within a sample. How about the enumeration of those populations, the actual counting of the cells? What are the issues?

Well, one of them is simply sample acquisition. One needs to be able to first able to identify the population and distinguish normal from abnormal. And this depends on the degree of immunophenotypic aberrancy, as we have described,

and the number of immunophenotype in the normal background population cells. It's the interplay between those two that, in large part, determines the sensitivity of the assay.

How many events define a cluster or a population in these types of analyses? In the literature, most people use somewhere between 10 and 50 events to do this. We typically use 10. In some cases, when there are fewer events, if the assay is clean and everything looks good, we can go lower than that. But something in this range is what's commonly used.

The enumeration itself requires the ability to completely discriminate the population from the normal precursors in order to accurately quantify it, which means that one needs a sufficiently informative immunophenotype to completely separate the two populations.

This includes the ability to separate the maturational expression that's present in many of these normal counterparts that are present in the sample, as well as occasionally within the leukemic

population itself.

The reproducibility of these enumerations, beyond just the ability to identify the population, then depends in large part on Poisson counting statistics, where the CV is the square root of N over N, so that 100 events in a given population, that one is looking at of MRD, will give a CV of about 10 percent at minimum.

So this means that if you want a sensitivity of .01 percent with a reproducible enumeration, with a CV of about 10 percent, you need to acquire about a million events, which is a fair number of cells to put through the instrument. And until recently, I would say beyond the capability routinely in many clinical laboratories.

Another issue related to enumeration is which denominator to use. This kind of plagues the field at the moment. There have been a number of different denominators that have been used in publications, ranging from the number of total nucleated cells in the sample, which is probably most comparable to what is commonly done by

morphology, to do this DNA binding dye, such as Syto16 or Draq5 are often used in order to exclude mature red cells that lack nuclei from the analysis if they haven't been removed by lysing or to remove the presence of platelet aggregates.

The concern of this denominator is that there may be underrepresentation of nucleated red cells because of lysing and washing. Our experience has been that that's not true, but it depends a bit on the processing procedure that's actually used.

Another denominator that's very common is white cells. This is probably the most commonly used by clinical laboratories, where the total number of leukocytes, or CD45-positive cells, including all the neoplastic cells in case they've lost CD45, are included as the denominator.

There is a variable degree of CD45 expression on normal immature erythroid cells in the marrow so that it results in some subjectivity to what's included or excluded. And if there is a lot of erythroid hyperplasia, one can overestimate

the level of involvement relative to what one sees with total nucleated cells.

The denominator that is currently used in the Children's Oncology Group ALL studies is mononuclear cells. Here, the granulocytic lineage is excluded by high side-scatter. And this is, in part, an attempt to mimic what happens when one Ficolls material, which had been used in many of the prior studies at the time these assays were set up. So much of the early MRD literature uses Ficoll material. In addition, because many of these samples are being shipped from long distances, there was concern about granulocytic degeneration in these shipped samples, and this removes that particular concern; so three different kinds of denominators being used.

This is an example of what these denominators might look like, by forward- and side-scatter. The nucleated red cells are colored in dark blue and the white cells are in gray.

One can draw a gate or box around the nucleated cells using Syto16. This would be a

total nucleated cell denominator. Just below that, one can use CD45 to identify just the leukocytes that are CD45 positive and exclude the nucleated red cells, or one can include the nucleated red cells, but exclude the granulocytes, as is shown in the bottom left. So these are each different ways of identifying the denominator.

I would point out that in the forward side-scatter plot, one needs to take care if one is going to use or include nucleated red cells in the denominator, that the instrument is set up appropriately so that they are actually included above the discriminator or threshold because they often have been compromised by specimen processing, particularly the lysing reagents, and have relatively low forward scatter. And this may result in some of the concerns that people have expressed about the loss of this population as a denominator.

What of the utility of including a DNA binding dye in these types of assays? Recently, we took a look at this because this question had come

up in some of our discussions. And so we took 100 samples of B-lineage ALL that were positive for MRD at day 29 and compared the MRD quantitation or remuneration in the presence of the DNA binding dye and in the absence or not using the DNA binding dye to find the denominator.

One can see that the effect here is relatively minimal, that the vast majority of samples actually -- there is good concordance in terms of remuneration. And in particular, on the bottom right-hand side, one can see that there are essentially no events near the therapeutic threshold that's being considered where this denominator has much impact.

Having said that, there probably are rare cases where this will be an issue, but in general, it seems like the use of DNA binding dye doesn't necessarily result in a different type of result from a denominator that does not use that.

Another issue in enumeration in bone marrow samples is hemodilution. Bone marrow is a semi-solid tissue. Absolute cell concentrations have no

actual biologic or physical meaning. The marrow aspiration itself is a traumatic procedure that introduces a variable amount of blood. And this tends to increase in amount as one does subsequent aspirations. So one should really use the first bone marrow aspirate for MRD detection.

For many samples, this turns out not to be a major problem, but in certain cases, such as hypocellular marrows, or patients with high white counts in the peripheral blood, or just poor quality aspirates, where one's aspirating, basically blood, it can be an issue. And at this point, there's no real accurate method for correction, although a normalization method has been suggested.

So what are the sources of variability for these flow cytometric MRD types of determinations? In terms of the population identification, either false-positive or false-negative results, these tend to lie in the areas of having an insufficiently informative combination of reagents to really allow identification of the leukemic

population and to separate it from normal. And this is I think a significant problem with some of the assays that have been published to date in terms of their ability to do low-level MRD detection.

Some of these assays in clinical laboratories are not really capable of being validated because of a lack of access to sufficient numbers and types of samples in order to do true assay validation. This poses some real problems in terms of the quality of these assays.

As I've mentioned, immunophenotypic shift can be a significant source of variability, depending on how one takes that into account. And using inexperienced interpreters that have had insufficient access to either normal material, regenerating marrows, or leukemic samples can result in false-positive or -negative results as well.

In terms of quantitation, the basic issues are that, if one acquires too few events, one cannot either accurately identify or enumerate MRD.

The effect of these varying denominators that we've described has some impact, although it probably has an effect in the range of about a twofold difference in estimation. Given that we're looking at a dynamic range, it's usually measured on logarithmic scales for MRD detection. This effect is probably not the major source of variability in these types of assays.

Sample degeneration certainly can be an issue in these types of assays, particularly in large oncology group studies, where they're shipped from long distances. And we have some experience with samples from Australia that suggest that at low levels of MRD, there may be false-negative discordances that arise due to sample shipping. And then hemodilution, depending on the degree, can have an impact.

So given all these caveats about the assay and the way in which the technology is used, what has the reproducibility of this assay demonstrated to date?

So there have been relatively little data

published on this, but the BFM group has produced some of this data in a nice publication a few years back, where the four labs performing flow cytometric MRD were assessed relative to each other, initially by the exchange of the list-mode file data, so just analyzing pre-collected data to see what the level of concordance might be. And this is what they demonstrated, that one can achieve reasonably good concordance in terms of just data analysis if one looks at the same set of data.

If one exchange is paired samples between two laboratories on the left, one can see that one can also achieve relatively good concordance. And if one creates artificially diluted samples, which are more stable and can be transported between multiple laboratories, one can also achieve a reasonable level of similarity or concordance in terms of the ability to identify MRD.

I want us to keep in mind, these are four laboratories that have spent a significant amount of time invested trying to standardize, relative to

each other, so that they get relatively comparable types of results. They're not just for laboratories out in the community that are trying to do these types of assays.

If one looks at how this affects the reproducibility at the various time points that were being measured, one can see that in the four different laboratories, on the top, that regardless of the particular time point at which the data was being measured, all four laboratories were showing reasonably good concordance. And this also translated into assignment into risk groups, low, medium, and high, on the bottom chart.

The COG experience, I would say, has been very similar. Mike Borowitz and I are the two laboratories involved in the pediatric ALL MRD work. And we've taken some efforts to standardize the assays used in our two laboratories. Between the two tubes and the assay that I previously showed, this is the reproducibility of quantitation between those two tubes, which is I think very high.

If one then looks at many thousands of patients, in this case close to 4,000 patients that were present on the last generation of the COG trials, one can see that at either day 29, day 15 in the bone marrow, or day in the peripheral blood, the rate of MRD positivity, at a level above .01 percent, is essentially identical between the two laboratories. And if one looks at the different levels of MRD, one can see that the range, whether it's between .01 to .1, .1 to 1, or greater than .1, is essentially the same between these two laboratories, in this case on the subset of data from the high-risk protocol.

This translates, then, into similarities in outcome. So this is the difference in outcomes between east coast and west coast laboratories for patients in the range of .01 to .1 percent, which is essentially identical, between .1 and 1 percent, which is again essentially identical, and including a relatively interesting shape to this particular outcome curve because this is the subset of patients that were given additional therapy, and

for patients with higher levels of disease, the slight difference being largely due to low numbers of samples.

So on COG, our experience has been that, at least between two laboratories, one can have a very high degree of concordance for the detection of MRD, even down to a level of .01 percent.

So in conclusion, I would say that flow cytometry is capable of accurate MRD assessment. However, it is an interpretive assay, and it's highly dependent upon the individuals performing the assay and the methodologies used. So it needs to be done by experienced individuals.

Its sensitivity is dependent upon the antibody combination that's used, how informative might that be; the number of cells that are being evaluated, a sufficient number need to be collected; and the time point that's being evaluated.

This type of technology I think will always be less sensitive than PCR, in part because, in principle, PCR can detect down to essentially a

single cell and amplify that signal. In order to identify a cluster of events to look at, we're talking about at least 10 cells that are present in that cluster. And that's one log difference right there, which I think is very similar to what Jacques has been showing.

At the current time, routinely, we believe that one can achieve a sensitivity of .01 percent for ALL at early time points, such as day 29, where the number of normal B cell precursors is lower. At other time points, where the number of hematogones or precursors is higher, the sensitivity will be variably reduced, depending on how informative the combination of reagents might be.

For AML at this time, a level of .1 percent is routinely achievable. And in a major subset of cases, one can go below that. But again, it depends on the phenotype and the panel of reagents used.

As I've shown, there can be pretty good reproducibility between different laboratories performing this assay, provided one takes the time

to try and standardize the assays. And this type of MRD detection does appear to correlate with clinical outcome. Thank you.

DR. BAGG: Thanks very much, Brent. Our next speaker, from Monza, is Gianni Cazzaniga, who's going to be talking more on a molecular-centric approach.

Presentation – Gianni Cazzaniga

DR. CAZZANIGA: Thank you. Thank you for the kind invitation. Although the title of my presentation was molecular monitoring of MRD, I've been asked to focus more on the comparison of flow and the PCR monitoring of MRD, according to a recent paper that we have done in our lab.

In talking about different techniques, I want to go back to some definition of prerequisites for reliable techniques, meaning that a list of topics that need to be addressed would first be the sensitivity, as already discussed, at least 10 to the minus 4, although it depends on the clinical question because not in all the cases can be requested such a high sensitivity.

Then the specificity to prevent false-positive results, the quantification, of course, because now we need to have a large dynamic range of quantification. And then also important is the stability of the marker, in particular, monitoring late time points for long-term studies. And then also important is the reproducibility between laboratories, in particular when we are talking about multicenter trials.

We already heard about the standardization and quality controls that are required for adding validation of the technique. And also important is to have the results in time for their clinical usefulness, which need to be rapid in most of the cases.

I want to stress again that when we plan MRD-based studies, important, independent of the technique, is to validate the MRD analysis according to the clinical setting and to the specific question to be addressed. So it's not possible to easily transfer one MRD setting to another without considering the therapeutic scheme,

the time points, the cutoff, and then the sensitivity, and also the lab expertise.

I would like to skip these lines because the definition of the biomarkers have been already discussed in the early morning. I just want to say again, after Jacques' talk, that in this presentation, we used the PCR guidelines defined in the EuroMRD group. And in particular, we have defined the so-called quantitative range, which is the lower limit of quantification of the assay, and the sensitive range, which is the lower limit of detection, where the problem for the PCR is the amplification from a non-specific evaluable region of the immunoglobulin T cell receptor gene rearrangement that could occur at very low dilution, affecting the sensitivity at very low levels.

I would like to go directly to the comparison of the RQ PCR and flow cytometry. It has been already discussed that the sensitivity is a bit higher for the RQ PCR compared to flow. And in both cases, the specificity is quite high. For

the PCR, it is clone- and patient-specific sequence, which is monitored.

A major difference is in the time of response, being quite slow for the PCR because we need to prepare everything. So it's very time consuming and labor challenging, the preparation of the probe, why is it more fast.

For the flow, we didn't discuss about the cost. Maybe we can open some discussion on that. And the standardization is particularly high for the PCR, and we are in the frame of the EuroMRD. And it's improving, as we heard before, for the flow, but still, there are limitations in the standardization. And the applicability is very high for both methods. And important is that the quantification is a bit different. It's based on the log-level reduction for the RQ PCR, and it is based on the percentage of cells for the flow.

So as Jacques mentioned in his presentation, there are several papers that describe -- that compare PCR versus flow, and this is an incomplete list of the papers. But most of them were in a

limited number of patients or in monocentric studies. And that's why we decided to compare, in our setting, the concordance and flow and PCR in a paper that's been recently accepted for publication in Haematologica.

In this paper that I will show you in detail, we analyzed more than 1,000 patients out of the almost 5,000 patients eligible and analyzed within the AIEOP-BFM ALL trial. And those patients are representative of this large group of patients. And in total, we analyzed 2,700 samples at different time points. All of them have been analyzed at day 33 and day 78, but also day 15. And there was no selection on the cases.

I would like to emphasize that this is the largest number of patients prospectively analyzed. And based on the routine application of both RQ PCR and flow, and a routine, independent application -- and this is a multicenter study, not a single center, with a centralized collection of samples in Italy, Germany, and Austria. And we compared different time points, but also, we have a

longitudinal monitoring along different time points for the same patients. Then we also did some comparison on the sample preparation for flow. And we also had the chance to look at the outcome, in particular, for the MRD discordant patient, which could be quite interesting.

So the methods are there are two PCR monitoring of immunoglobulin T cell receptor gene rearrangements, which is based of course on DNA from mononuclear cells, while the flow cytometry is based on the same sample, but on nucleated cells before separation of mononuclear cells.

The analysis was done according to EuroMRD guidelines for the PCR, and flow was based on the method described, cited in the previous speaker, in the AIEOP-BFM group. The sensitivity was at least 10 to the minus 4, as requested by the protocol, for the RQ PCR, and 300,000 events were acquired for the flow, allowing to reach a sensitivity of 10 to the minus 4 in most of the cases.

There are several variables to be considered at different time points. As already mentioned,

there is a problem of normal regeneration, which is different at different time points, as indicated here.

The other point is that, hopefully, the MRD is decreasing during time, so it's quite high at day 15 and then is reducing during time, challenging the two methods because of sensitivity -- the requirement of sensitivity is increasing.

The other problem that has been already discussed is the shift that could occur by monitoring flow cytometry, as described in previous papers, that need to be taken into account.

So looking at the concordance, at different time points, the overall concordance is 80 percent, considered in the total number of samples analyzed. But the concordance is quite valuable at different time points, being quite high at day 15 and day 78.

But the lowest concordance is at day 33. And what was indicated, the major discordances, which are positive by PCR over the cutoff of 10 to the minus 4 and negative by flow with the same

cutoff.

If you consider the flow cytometry sensitivity by assuming that the PCR is the gold standard and the sensitivity is decreasing during time, being 87 percent at day 15 -- and this is the number that considers the positive by flow over the total number of positive by PCR. And it's going down because the MRD is decreasing, and the specificity is increasing, where the specificity is defined as the capability of flow, to call negative, a sample which is negative also by PCR.

So by looking at the direct comparison of MRD at the different time points, it's quite clear, as already highlighted by Jacques, that most of the discrepant cases are negative by flow or below the cutoff 10 to the minus 4 and positive at a different level by PCR at the different time points.

This is even clearer if you look at the concordant positive in red, or orange here, or the discordant PCR positive and flow negative at the different time points. As you can see, when you

have discordant cases, PCR positive and flow negative, those are at least one log lower in terms of PCR positivity compared to concordant.

This is even more clear if you consider that the discordant PCR-positive and flow-negative cases are not -- so the reproducibility of the PCR at this level is not that high, because if you consider positive sample in two out of two markers or in one out of two markers by PCR, this percentage is decreasing during time; so two out of two, while the one out of two is increasing during time, indicating that we are at the limit of the sensitivity of the methods.

Here, also we look at the different source of the sample for the flow, demonstrating that there is no major difference if we use nucleated cells or mononuclear cells for the quantification, although this is done in a laboratory in a limited number of patients.

So at the end, we also look at the outcome. In particular, we focus on day 33, where the discordances were higher and, as expected, maybe

the double negative are the very good patients, and the double positive were the poorest patients. And interestingly, the discordant cases, in both sense, PCR-negative/flow-positive, or PCR-positive and flow-negative, gave an intermediate and very similar outcome.

More in detail; if you look at different subgroups of patients here, they are quite similar. The curves are not statistically different, although there is a tendency for the lowest values to be better than the others.

So in conclusion, the overall concordance was around 80 percent and disconcordances were mostly due to difference in sensitivity of the two methods. And they're more frequently due to flow cytometry negative in cases, PCR positive, at lowest level of MRD. There is no major impact of the source of cells.

Also, we found with time point, this dependent concordance, which could be due to different reasons, so the regenerating cell background, definitive modulation, different tumor

burden, and of course the sensitivity. And it's important to realize that the two methods measured different cellular targets. One is DNA based and the other is protein based. So the complete quantitative concordance is also unlikely.

Concerning the outcome, we saw that the double positive and double negative are the extreme of the curves. And the intermediate outcome, which has been observed in both discordant groups, may represent the effect of intermediate or very low levels of MRD at the limit of the sensitivity of both methods.

These results could suggest also a potential complimentary role of the technologies for the improvement of treatment tailoring. And I have to mention that this is already the case because, as you are aware, by the publication of Basso and colleagues, by flow at day 15, they demonstrated that flow cytometry at this time point is able to stratify patients to three different risk groups. And the same is used since a long time for PCR at different time points, day 33 and day 78. And

also, the choice of the method depends on several reasons, including the expertise, the resources, and the design of the trial.

In the last slide, just to show you an example of the combined use of flow and PCR, I would like to show you the results of the comparison of the stratification by either one or two sensitive markers by PCR in the closed AIEOP-BFM ALL 2000.

By using two markers in that protocol, we were able to stratify 77 percent of the patients, so 33 percent of the patients were not stratifiable by PCR MRD. But by using only one sensitive marker, this percentage decreases to 12 percent. And this is possible because if you compare the use of one sensitive marker -- or two sensitive markers, the outcome is not that different. And that's the reason why, in the new protocol, only one marker is required.

The other point, as I mentioned, is the availability of the flow at day 15, that can be used to stratify the patient. So in the new

protocol that's already open, the 12 percent of patients that are not stratifiable by PCR MRD can be stratified according to flow at day 15, with this cutoff indicated here. And on top of that, the cases that are high risk by flow are going directly into the high risk in the ongoing protocol.

So I will stop here. And I would like to thank all the people that collaborated to this, in particular Giuseppe Gaipa from our lab, who is the flow expert for us. Thank you.

Clarifying Questions to Presenters

Group Discussion

DR. REAMAN: We could probably open these up for a few minutes of questions.

Gianni, can I just ask, I understand that the PCR was done according to standardized methodology, but with the PCR, was there standardization there as well? And were there any efforts to evaluate the concordance between labs using test cases prior to your study or as part of your study?

DR. CAZZANIGA: You mean by PCR?

DR. REAMAN: No, no. By flow.

DR. CAZZANIGA: By flow.

DR. REAMAN: I'm sorry.

DR. CAZZANIGA: Yes. But by flow, the standardization is the one that has been performed in the paper by Dworzak in 2008. And in this paper, we described that, as I showed you in the previous presentation, in different centers, there was concordance and there was exchange of samples and exchange also of technicians going in the different centers. And this is a continuous process that is ongoing. Yes.

DR. WILLIS: I'm Tom Willis from Sequenta. So it won't surprise anybody who knows Sequenta, including some of our collaborators here, that my question relates to the potential of a new platform for MRD detection, which attempts to combine the best features of PCR in terms of its sensitivity and specificity, and the ease of use and general universal assay of a flow cytometry assay, by using a PCR assay of the immuno-cell receptor

rearrangement, but a sequencing-based readout rather than an individual PCR reaction for each patient, which attempts to, again, combine that sensitivity you get with the PCR with a more generally applicable technique.

I guess I would ask for the platform presenters' comments on the potential of those kinds of techniques.

DR. VAN DONGEN: So the question is about whether the high-throughput sequencing could replace the current PCR technology. Theoretically, this could be done. There would even be some advantages to use that technology because you identify each individual sequence.

How this would work out in terms of precise quantitation, you're only quantifying within the detected rearrangements. And while you do it within PCR, you also quantify against your total background of DNA.

Now, you're only picking out your rearrangements from your B cells. So you quantify within your B cell population. So then you still

have to determine what is the size of your normal precursor B cells and B cells in your total population that you have analyzed.

So that's not taken into account in all the currently methodologies with high-throughput sequencing. But as such, I think it's an interesting approach. It's still very expensive. At least, I am not aware of large-scale studies. There is some interest now in Europe to use it as a new approach.

The main advantage could be early detection of CYP confirmation and preventing false-negative results in the PCR methodology due to changes in the rearrangement pattern.

DR. REAMAN: Dr. Cazzaniga?

DR. CAZZANIGA: Yes. I will agree with Jacques and just would like to make another comment that was related to the presentation of Andrea, showing the heterogeneity of the ALL. So can we forecast this to complicate things is that we are going to see also small subclones, a diagnosis that then needs to be discussed and to interpret in the

appropriate way. We don't know so far what will be there.

DR. WOOD: And I would agree as well. We have some feasibility data that will be published in Science Translational Medicine this year for T-ALL that shows exactly this.

DR. WILLIS: I'd just say, in response to Jacques' comment about expense, there's one thing you can bet on in the world today, is the fact that DNA sequencing gets cheaper every year. So just wait, and it will be cheap enough.

DR. REAMAN: We'll all hold our breath.

(Laughter.)

DR. REAMAN: So just maybe one more question. We can have some discussion during our lunch break. And then I think when we actually address questions after lunch, we can have more discussion about the specifics of the technology platform.

So go ahead.

DR. JENSEN: I'm Leeann Jensen from the National Cancer Institute. And I wanted to ask

Dr. van Dongen in particular, and any of the other speakers from this morning, the data that you showed about the standardization of the normal bone-marrow regeneration versus the treated bone marrow; is any of that available? Is any of that data from adult patients or is it all pediatrics?

I've noticed a general focus on pediatric data, because that's where most of the work's been done, but I'm wondering if there might be differences in adults to children.

DR. VAN DONGEN: So the data that I showed was indeed pediatrics. However, we have done this analysis as well in adult patients. And the pattern of regeneration is comparable in that sense. And that's what you heard already from these other two speakers as well.

It is dependent on the time point. It's also dependent on the preceding treatment, in fact, the massiveness of the regeneration, that's really different per protocol. We evaluate three different treatment protocols using exactly the same time points, the three same time points. And

we got a different regeneration patterns, per protocol. And our impression is that the intensity of the preceding protocol, the preceding treatment block, is determining the amount of regeneration.

So the harder the preceding block, the more massive regeneration it is. And massiveness of regeneration means more immature precursor B cells. And they are the contaminating problem for recognizing the malignant precursor B cells.

So again, I think you can extrapolate it to the adult setting completely.

DR. REAMAN: We'll break for lunch. I believe there are food items that are available out in the lobby. You can find someone who's an FDA employee that could escort you to the cafeteria. But otherwise, we'll resume the workshop at 1:30. Thanks.

Whereupon, at 12:32 p.m., a luncheon recess was taken.)

A F T E R N O O N S E S S I O N

(1:35 p.m.)

Questions to Workshop Participants, Decisions and Summary

(Audio gap – approximately two minutes.)

DR. HUNGER: I would argue that the answer is yes. I think large studies on both sides of the Atlantic have clearly established that early response is measured by MRD, is the single most important prognostic factor that's predictive of outcome in ALL.

There certainly are nuances as to the technique and the level, but I think, with both techniques, patients -- or with all techniques, patients who have a reduction to less than 10 to the minus 4 at the end of a month of therapy, whether it's newly-diagnosed patients or relapsed patients, have a markedly better outcome than those who do not have such a production. And in multivariable analyses, it tends to be the single most important prognostic factor.

So to me, I think that that makes an

argument that this is a surrogate marker. Obviously, the details of study design and other things I think are very important. But that would be the point I would start discussions from, personally.

DR. BIONDI: Just to make a comment to support what Steve said, I think, with respect to the subheading A, if we really would be in the position to make a meta-analysis of the data so far, I think it could probably get a conclusion to support this. We are talking about productivity of MRD as a surrogate for EFS and survival. So we're not talking about activity long term. We're just talking on what we have seen this morning.

I would say that all the studies in adults and children as well fully support this conclusion. So I would just state that probably it would be nice to make a sort of meta-analysis of the data and to conclude that.

DR. REAMAN: So can we have some discussion about how we might actually accomplish a meta-analysis?

DR. GOKBUGET: As I mentioned in my talk, for the adult ALL in Europe, we already planned a meta-analysis. And I think the most important point, of course, is to very clearly define which type of patients will be included, which data collected, which method, for example, should be floated or combined with PCR data? Should PCR-positive ALL be included, and so on.

So maybe we could exchange the definition of these meta-analyses for pediatric patients and adult patients, then later can be compared.

DR. REAMAN: So would you see actually combining adult and pediatric patients, or actually doing separate meta-analyses?

And I might just call on Wendy Stock, since most of what we've heard about MRD in adults, at least in the formal presentations this morning, have been the European experience.

Can you just give us a little bit of information as to where this stands in the U.S. as far as it being used for stratification on frontline treatment studies and the amount of data

that might be available, willing to be shared in the meta-analysis project?

DR. STOCK: So we haven't traditionally done too much frontline monitoring of minimal residual disease in the adult cooperative groups. In CALGB, our most recent study that we completed, which was 101.02, which was a study of all adults with ALL, we intercalated a module of alemtuzumab therapy. We do have data on about 70 patients, prospective monitoring, using PCR-based MRD analysis for outcome. We don't have all of the data analyzed, but that would be available.

We also have data in Ph-positive ALL. We have about 70 patients there, where we've prospectively monitored using PCR on a frontline study that is just going forward with a manuscript now. So those data are available.

In addition, our current frontline intergroup study, 104.03, we do have prospective PCR monitoring. Gerard Lozanski from Ohio State and my lab in Chicago have been doing both flow- and MRD-based analysis at serial time points, but

we don't have the clinical follow-up yet on those data.

Then we have a planned intergroup study that is going to be led by ECOG, which we hope will get approved where we're going -- where MRD, of course -- and that's one of the major driving forces for actually pushing this now, is that we're hoping to intercalate an antibody, blinatumomab, into frontline therapy. And minimal residual disease will be very important, if not the most important, parameter of response assessment.

So I think we have some patients. We don't have huge numbers, but we do have serial quantitative PCR monitoring on about 150 patients, with different types of ALL.

DR. REAMAN: So I get the sense that there is interest in at least the concept of considering a meta-analysis.

And go ahead, Grazia.

DR. VALSECCHI: Yes. Before somebody mentioned that the meta-analysis is based on advanced colorectal cancer, the meta-analysis

study, which the hazard ratio is plotted; we saw them before. So doing this meta-analysis to assess frontline protocols, we are talking, it would mean to put together frontline protocols, first of all, that are randomized with some treatment during induction, if we are looking at MRD at early time points; so talking about induction, end induction, and the threshold, the correct threshold.

So it would mean to pull data from trials that randomized some drug during induction, also different class of drugs, where although there are at least two trials that randomize dexa versus pred, so steroids. But I don't know whether there are others.

At least one of the trials didn't show a difference in activity, but it may be that looking more deeply into different cut points, these different thresholds would -- is this the sense of what you are thinking about?

DR. REAMAN: I'm just asking. So this is developed for discussion purposes. And in reality, I'm not sure that there are that many randomized

studies which have actually asked randomized questions in induction. It's pretty standard, has been standard, fortunately or unfortunately -- fortunately for children -- for a number of years.

I mean, there are certainly interventions based on risk, but I'm not sure that there would be many randomized questions in induction that would make this a burning issue for actually doing a meta-analysis in the newly-diagnosed setting. But it's out there for input from the investigator community.

Mark?

DR. ROTHMANN: Yes. I guess I was going to say some of the same things. I guess from the presentations, it's quite clear that baseline minimal residual disease is prognostic. I don't think we need to go much further on that.

But if you do want to know if it's a surrogate, you have to see if treatment effects on that endpoint, which apparently involves some sort of change in minimal residual disease, corresponds

to treatment effect on whatever clinical endpoint you're interested in.

Another thing I heard this morning, I forget from who or if it was from multiple people, was this talk that, perhaps it's in high-risk subjects who have a larger amount of minimal residual disease at baseline, where this could be a surrogate.

That's important. If subjects that have very little minimal residual disease aren't going to have much of a change in minimal residual disease, then, for them, it might not be an appropriate surrogate. And I think this is even related to what I believe I heard from Dr. Biondi -- apology if I mispronounce your last name -- in his presentation, where I heard him say that minimal residual disease was, in essence, predictive, that the treatment choice is dependent upon the minimal residual disease.

When that's the case, you're probably not looking at something that would be broadly a surrogate because two different treatments, it

would seem, would have the same order of effect on minimal residual disease, irrespective of the baseline values. But if one of them is preferred over another for subjects that have low minimal residual disease at baseline, which could be due to off-target effects, then you might not be looking at what would be a uniform surrogate for all subjects.

DR. VALSECCHI: Sorry. Just to add, within high-risk groups, since high risk is usually defined also based on a high level of MRD after induction -- and so patients are set to the high-risk group and are treated more intensively -- then within that high-risk group, treatment is uniform, in a way.

I mean, MRD is used to intensify therapy. Nonetheless, it stays an important prognostic factor, nonetheless. So it means the treatment wasn't able to cancel, let's say, the negativity. And then within those subgroups, it may be looked for in a surrogate. I don't know.

Is this what you meant?

DR. ROTHMANN: Well, I understand that, at the moment, all we're seeing is prognostic. But it may very well be the case that the endpoint will have large effects -- correspond to large effects on clinical outcome, perhaps for subjects who are high risk. But for subjects who have very small minimal residual disease, you may not get a relationship like that.

DR. VAN DONGEN: Coming back to this question, also in relation to the previous speaker, the question was also raising the critical threshold value because that's what it is all about. And I agree, for a high risk, it's relatively straightforward. If you now look to standard-risk groups in the different MRD-based studies, that's very variable. And it seems that we have used the same threshold -- and perhaps this is not the case, because for those data that are really compared data -- and Gianni showed the recent data that will be published in Haematologica, showing that flow and PCR are not measuring exactly the same.

So whereas I fully agree that it is obvious that there is predictive value, firmly, there's no doubt there, the precise risk of definition that you might be willing to use, also for starting new treatments -- or, let's see, one group you might like to use a new drug, and another group, you might not. I can understand, particularly the relatively large medium-risk group, that's where you'd like to change therapy because that's where there are relapses that you cannot predict. But then you have to be sure that your medium-risk group is defined very accurately. And that means we have to understand thresholds.

So far, what I've deduced from the literature, and what I've also seen in this meeting, is that the cutoff between the medium-risk group and the standard-risk group is not yet defined. That has to be done much sharper.

Whether this can be solved as a meta-analysis, then you have to do a meta-analysis separate for PCR and flow. But what we'd like to do here -- I mean, we like to understand how flow

can potentially even replace PCR. How far are we there? That's a discussion for the next part, but that plays a role in this question as well, because it's about the risk of definition. How sharp are the risk groups?

I think that's still different between the different treatment groups and the different technologies.

DR. HUNGER: This comment may bleed into the second question, but I guess I'm thinking we're having this discussion mainly in the context of, should this response characteristic be used in the drug approval process?

DR. VAN DONGEN: Yes.

DR. HUNGER: The drug approval process is most likely to occur in relapse patients, I believe, not in some of what we're thinking about in terms of optimizing treatments for patients or testing variations on standard drugs. But it would presumably be patients randomized to get or not get a new agent, or to get one of two different new agents. And I think, then, you're really going to

be focused on either the relapsed population, where the number of patients may be small and the patients are at very high risk of failure, or a subset of newly-diagnosed patients who have been culled out based on a very high risk of treatment failure.

So I'm not sure some of the nuances that may be very important about identifying the optimal method and optimal cut point for deciding whether patients are absolutely cured or not is as important in the context of accelerated approval of a new drug or for a new indication.

DR. REAMAN: That's exactly the point. So I think, to sort of frame this, where we would use this surrogate, would we choose to actually use it, would be in the setting of evaluating new drugs, which, as you mentioned, are generally in the treatment refractory-resistant population or the relapsed population.

I certainly agree, Jacques, with the need to define what is a critical threshold, depending on how that threshold is determined. But I think we

also have to consider -- and I'm going to move into the third question, also.

If we're utilizing MRD as a risk stratification criterion and actually adjusting therapy based on that risk factor, then can we, or how do we utilize the same feature as a potential surrogate endpoint? And how do we get around that? Or do we have to identify high-risk groups using criteria other than MRD at the end of induction or at some early time point?

DR. GOKBUGET: I would like to argue in favor of our strategy to test new drugs in patients with molecular failure, because most of the new drugs coming up are somehow targeted drugs. And molecular-failure patients are those who do not respond to conventional chemotherapy, so they are in need of something alternative. And they can be easily defined.

For example, in our new trial, we will have the option at the time point, after consolidation, to draw these patients out of the first-line treatment and to enter them into experimental

trials. And I think this is a very good group if you use them for evaluation of new drugs. But at the same time, you need then to accept that rendering them MRD negative would be the endpoint of this trial.

DR. REAMAN: So can you just define molecular-failure patients, failures at end induction or failures at any time along the treatment course? And would it require longitudinal surveillance of MRD and MRD monitoring?

DR. GOKBUGET: I think, to make things simpler, the minimum would be to evaluate MRD in the beginning of the treatment, either first-line or relapse treatment. In pediatric, it's a standardized relapse treatment, which is also effective, but after a certain duration of chemotherapy. For example, in our trial, it's consolidation 1. Then you don't get additional molecular remissions, and that would be the time point where these patients could be transferred into an experimental trial.

DR. REAMAN: Dr. Whitlock?

DR. WHITLOCK: Similarly, in plans for the upcoming COG relapse study for high-risk or early-relapse patients, we are proposing, after delivery re-induction chemotherapy to evaluate MRD status. And because of the published data about the very poor outcomes of patients who are in a morphologic remission, but are MRD positive, to provide those patients with additional therapy, they would be a molecular failure according to your definition before they go to transplantation. So again, a very similar strategy.

DR. BOROWITZ: Yes. I think that everybody sort of wants to do the same thing, but I guess the question that we really have to ask ourselves is, do we have to answer the surrogate question, the validity of the surrogate question, in every one of these different scenarios before we can think about doing this?

You presented some nice data, which are the first that I've seen -- and I have to examine them a little more -- which suggests that if you take

somebody from an MRD-positive state to an MRD-negative state, their outcome is better, which is what you'd really want to know. But at least that's the first data that I'm aware of that actually showed that definitively. We always assume that that's the case.

So I don't know if we're going to focus on this kind of approach, which might be a valid one. We really have to make sure that we know that MRD negativity -- MRD is not only a criterion for trial entry, but also a suitable criteria for response. And I don't know if in aggregate, there are enough data out there to justify it. Your data does speak to it, but I don't know if that's enough.

DR. REAMAN: I think you raise a good point. And I'm not aware that there's an abundance of data out there, either. And I think it would be required to evaluate the effect of an intervention or a new drug on the surrogate, as well as the final outcome endpoint. And that's a piece that's missing here.

I think it's something that could probably

be done simultaneously, as we evaluate a new agent within this setting of patients with molecular failure or early relapse.

DR. BOROWITZ: But that doesn't solve the problem of accelerated approval in that context.

DR. REAMAN: Well, it does if you could -- well, you'd have to follow the study through to make sure that the real endpoint, that's the gold standard for approval, overall survival, event-free survival, is also impacted by the drug.

DR. SRIDHARA: I think we have to get the terminology of it straight. We don't know that it is a surrogate yet. If we already knew about that it was a surrogate, then we wouldn't be having this discussion. So currently, it's a biomarker. It's an intermediate endpoint. And can this be used for drug approval is the question.

So unless we have good data from randomized studies, where you can show that a difference in this endpoint will result in a difference in a clinical benefit endpoint, such as even event-free survival or survival, it's not there yet.

So if it's not there, then you need to conduct studies where you measure this as well as you collect the survival. And so that's where I think we are. This is an intermediate endpoint, as has been looked at. And, yes, it's a good prognostic factor, but that doesn't mean that it can be used as a surrogate endpoint.

This is Dr. Sridhara from the Division of Biometrics.

DR. HUNGER: I think one of the problems we run into is the diversity of later treatment. An advantage of a surrogate endpoint such as MRD is, for example, you could conduct a randomized intervention. And we have a trial in development. I won't mention the drug's name, but we're planning to give patients induction therapy with relapse ALL induction therapy, and randomize them to receive backbone alone or backbone plus a new agent.

At the end of that first month of therapy, you can assess complete response rate by standard criteria, and you can assess MRD response. And you can get a readout, at least on the efficacy of that

drug in putting a patient into remission, which is relatively clean, so to speak.

I think the problem then becomes, what happens after that. And when therapy is not uniform after that, the eventual overall survival or event-free survival outcome may be clouded by the fact that certain patients had a bone marrow transplant, certain patients didn't.

I think those are the challenges you face, particularly when we deal with relapsed leukemia patients or very high-risk leukemia patients, where an ultimate goal of treatment is that a lot of them undergo some sort of stem cell transplant procedure, where many patients die of toxicity. That probably doesn't have to do with anything about their leukemia response. And patients may be cured for reasons that have nothing to do with their response to the drugs they got early on.

So to me, that's somewhat of the dilemma of the very high-risk population.

DR. REAMAN: But it's not totally different from other situations in which a drug is approved

based on overall survival, because there are situations where there are recurrences, and relapses, and changes in therapy that impact the overall survival, but may not really be related to the efficacy of the drug in question.

So again, I think -- and that's one of the situations that can be sorted out with the accelerated and regular approval process, using a surrogate for the accelerated approval and then with longer follow-up.

I think the other difficulty -- and I fully agree, we have a way to go with respect to the process that's needed to demonstrate or to prove surrogacy here. And we're sort of hamstrung by the fact that we have a very small number of patients on randomized trials in the relapse setting.

Going forward, I think we can certainly work to change that. I think it's going to require cooperation internationally, cooperation between adult and pediatric oncology. But right now, I don't think we would have, in the relapsed setting, a sizeable number of patients or a sizeable number

of studies to actually even do a meta-analysis in the relapse setting, looking at end induction or early MRD, unless people feel differently. But it just doesn't seem like that number is very robust to me.

DR. PAZDUR: One of the things that I want to point out for accelerated approval, it's not that we have to establish that this is an established surrogacy. Regulations clearly state that it's reasonably likely to predict clinical benefit.

So I don't want to have you get hung up on the issue that this has met the Prentice criteria or something like that, that this is an established surrogate. It's particularly brought out in the regulations that it is reasonably likely to predict clinical benefit, and then that would be verified by studies that would come afterwards.

The other point of view is, we're having a discussion here of an endpoint as if it was in absolute-value bracket here and not talking about the magnitude of benefit, because obviously, if

you're talking about a drug that would have a significant number of patients receiving, achieving this minimal residual disease status, 50 percent, 80 percent, 90 percent, that's a much different drug than if 5 percent of the patients have it.

So when you're talking about a drug, particularly in a refractory disease setting, one would have to think of also the magnitude, just like when we talk about response rates in refractory disease populations in solid tumors, for example. There's a big difference between a drug that has an 80 percent CR rate versus a drug that has a 10 percent CR rate. Okay?

This is a point that I think needs some discussion here also, magnitude of change that one would see in a refractory disease population, that would perhaps translate into an impact on event-free survival.

DR. REAMAN: I think you're absolutely right.

DR. PAZDUR: I'm just looking for a number.

DR. REAMAN: We don't have a good number. I

mean, I think with conventional re-induction therapy, utilizing like the COG platform or backbone, only 25 percent of patients following block 1 are MRD negative.

So if you gave that backbone plus epratuzumab, plus some other drug, and it was 50 percent, 75 percent, 80 percent, then I think there, the kind of situations in which we would have objective criterion and effective data in which we could put forward the consideration of a likely surrogate for clinical benefit.

Did you have a question or a comment?

DR. LINK: I was just going to respond to Rick that these patients aren't refractory in the sense of -- in other words, the induction rate, depending on which population you choose, is actually pretty good using standard criteria. So it's a little bit different.

DR. PAZDUR: Okay. Different.

DR. REAMAN: But morphologic re-induction --

DR. LINK: That's what I'm talking about.

DR. REAMAN: But we're now sort of looking

at a new level of sensitivity or maybe a new way of defining response, which would be MRD rather than CR rate and finding less than 5 percent blasts in the marrow.

DR. LINK: No. I was just commenting. But ultimately it's got to be tied to something that everybody knows is associated with an improved, event-free survival, which is morphologic remission, not MRD remission. I assume it is.

DR. WHITLOCK: So I would agree. The surrogate question, we may not be ready to address. But clearly, MRD is just a refinement of looking at complete response.

I think we would all agree that that's a clinically relevant endpoint and that MRD is a refinement of that. And so in that sense, it seems intuitive that this is a clinically relevant and important evaluation. We just don't know what the precise cutoff values are at the present time.

DR. REAMAN: So we're sort of going out of order of the questions and maybe not following them at all.

(Laughter.)

DR. REAMAN: But I think we have a string going here. So we've moved out of the newly-diagnosed patients. We've sort of focused on relapsed or molecular-failure patients, potentially high-risk patients. But we still have the issue that Dr. van Dongen mentioned. And that is threshold level or critical threshold level.

So is there some discussion about what might be an acceptable threshold level in which to strive for, depending on which of these patient populations we would go forward with?

DR. WHITLOCK: I think the bulk of data that are available today suggests it's 10 to the minus 4.

DR. GOKBUGET: I would also support this. For the panitumumab trial, which is ongoing in Europe -- I have mentioned it -- it's done in MRD-positive ALL. We have selected the threshold 10 to the minus 3. And this is a very bad threshold because it's short before clinical relapse, so sometimes it's really no time to even

organize a trial. So 10 to the minus 4 is, in my point of view, very good.

DR. BIONDI: Although, if I go back to the original paper by the BFM on the S2, that we produced together with the Berlin group, it was 10 to the minus 3. The cutoff really discriminates those patients within the S2 category that did well, as compared to after the first course of re-induction.

DR. HUNGER: I think it might have a lot to do with trial design and numbers. We know that MRD is a continuous variable. The more you got, the worse you do. And we know you can dichotomize it at numerous cut points. And if you dichotomize it at 10 to the minus 3, you're going to have fewer positive patients, but they're going to do worse than if you dichotomize it at 10 to the minus 4.

So it may very much depend on the treatment effect you're looking for and the sample size you have available to ask a question.

I think any cut point is relatively logical and clearly predicts outcome. Whether it predicts

therapeutic benefit, I think is a different question, but it's very clear that it predicts outcome.

DR. LINK: One problem with choosing a value is that technology is going to change. In other words, within a couple years, you'll have something that's better. And then the question is whether it'll be more discriminating or not. So the question is that you ought to accept the principle and then you can decide -- you know, the cut points are going to be determined by what technology is available, the best technology that's available.

DR. REAMAN: But I think, in order to really define the principle that you're accepting, you have to start somewhere and define it.

DR. LINK: Right. I think, just like MRD is a continuous variable, I think the technology is also a continuous variable that's going to be changing and evolving over time. And you're right.

DR. HUNGER: Unless you're willing to bleed somebody dry, there are some limits to technology. I mean, you have to decide how many cells you're

willing to sample, and routine clinical practice and. At least today, the way you sample cells by doing a bone marrow aspiration and taking out some number of CCs of marrow, it gets hard to capture more than, at least as far as I know, somewhere between 100,000 and a million maximum possible readouts.

So if you have a technology that can detect 1 in 10 million cells, you have to get 10 million cells first for it to be relevant.

DR. VAN DONGEN: And also the technology plays a role in terms of whether you can amplify a single piece of DNA or whether you have to detect a group of cells within a quite narrow definition, most likely being at least 20 cells.

If you go for at least 20 cells and acquire, theoretically, 2 million cells, you are not down below 10 to the minus 5, although 2 million sounds a lot. Then PCR logically with the same amount of cells will be more accurate.

So the technology determines also, at least in part, the number of cells that you have to

analyze. And again, we talk now about cutoffs, but a cutoff is also dependent, most likely not only depending on the treating protocol, the design, but also the timing. So timing plays a role. So we will have a lot of variables.

But the technology, that's clear. The moment you talk about cells that are analyzed, it's not one cell. You talk about at least 20 cells to have a clear positivity; maybe 10 cells if you have very well-defined markers. And that's different from DNA, also different from RNA, where you have multiple copies. And we are still not aware of what we are actually measuring because you have thousands of copies per cell.

DR. REAMAN: So we're actually getting into the issues related to technology, which I was hoping we could hold off, but I knew they were going to be very closely connected and clearly interrelated.

So maybe we can't come up with a specific critical value. But I would like to just continue the discussion a bit in the relapsed patient group,

how we would actually qualify early MRD as a response biomarker and ultimately be able to use it as a surrogate for clinical benefit.

Jim?

DR. WHITLOCK: So currently, we look at the percentage of patients who enter a morphologic CR, so it would seem to me logical to say that we look at the percentage of patients who achieve a predefined molecular CR or level of MRD, as whether a drug is effective or not.

Particularly in the context of the randomized strategy that you're talking about, which we're planning to evaluate in COG, to have a platform of effective chemotherapy, and then in a randomized way look at the incremental benefit of a novel agent added to that, then one would see just a change in the percentage of patients who achieve a level of MRD.

DR. REAMAN: Would you see doing this with a time point after the first block of therapy, after a second block of therapy?

DR. WHITLOCK: Yes.

DR. REAMAN: Third block of therapy?

DR. WHITLOCK: We're planning to do it after each of the three blocks of therapy, before patients are then eligible to go to transplantation. Then additionally, patients who remain MRD positive after receiving this therapy would not be eligible to go to transplantation based on existing data about their poor outcomes. And we would like to evaluate the strategy of introducing a novel agent for these molecular failures and see if that can bridge them to then go to a transplant if they become MRD negative.

So I think multiple evaluations are appropriate if we're really going to be able to identify the best future point for evaluation.

DR. HUNGER: Maybe I'd like to at least get a scenario where we could be concrete. And this is a trial design we're considering. So patients with relapsed ALL were in a group such that almost all of them go into morphologic remission, but only about 50 percent of them go into an MRD-negative state.

So if you gave these patients a month of chemotherapy with or without a new drug on top of that, what if, at the end of that first month, the standard arm without the new agent had a 50 percent MRD-positive rate and the standard drug plus the new agent had a 25 percent MRD-positive rate? Essentially, a hazard ratio of .5?

Somebody better than I would do the numbers and tell you how many patients that would require, et cetera, and the power. But just throwing that out as a straw man, would that be something that people would think merits consideration of accelerated approval, a design such as that, with follow-up to see if there's an ultimate benefit?

DR. REAMAN: I would think so. I mean, it's a 50 percent -- it's a doubling of the percentage of patients who don't have a detectable disease at a time point. The question of feasibility, given that, I'll leave to our statistical colleagues. But I think that's the kind of robust evidence of efficacy or response that Dr. Pazdur alluded to earlier, at least for an accelerated approval

consideration.

DR. PAZDUR: One of the things that has haunted us, as far as taking complete response rates, is when we talk about responses, we're usually talking about what is the duration of that response, and then it becomes confounded by subsequent therapy, patients notoriously going onto transplant, et cetera. And we really then lack that. And that has really made the, I think, evaluation of drugs in leukemia in general difficult, this "what is the duration?" And that's why we were looking at, is this a way of perhaps getting around that issue, by looking at MRD here?

DR. HUNGER: Since MRD is generally used as a criteria for who goes to transplant, in the scenario I just gave you, if the end re-induction MRD level is used as a criteria for transplant, only half as many patients will undergo transplant.

We could argue whether that's a good or a bad thing. Many of us in the room would feel that that's a good thing, to avoid a treatment modality with a known risk of toxic mortality.

So I think, yes, the issue of transplant confounds it, but also the achieving of better response rates may prevent the need, potentially, for transplantation.

DR. WILLIS: One thing I'm just confused by this conversation is, is Bcr-Abl and CML not the right framework for drawing analogies for this?

I guess, to what degree does the nice example we heard this morning of Bcr-Abl as a validated surrogate endpoint that was used to approve drugs not a basis for having this discussion, in answering, for instance, Dr. Hunger's question as posed?

DR. REAMAN: It's much easier. It's specific. And we're talking about a much more heterogeneous population --

DR. FARRELL: And disease.

DR. REAMAN: -- and disease. So that's the quandary, I think, or part of the quandary in which we find ourselves.

DR. WILLIS: It's the connection of the mechanism of action to the drugs that were

approved, to the biomarker that was the added thing that gave that more clarity than we have in this case. Is that the --

DR. REAMAN: Right.

DR. HUNGER: The standard-of-care treatment is monotherapy. I mean, I think it's so much cleaner. It's like everything related to Bcr-Abl. It's a great model that almost nothing recapitulates because it's monotherapy, which you don't treat any other -- or few other cancers with monotherapy.

DR. DEISSEROTH: The other thing that might distinguish and separate the CML experience from the ALL experience in terms of talking of MRD, the MRD endpoint in CML, at the time that it was proposed as a regulatory endpoint, had already been shown to correlate with the outcome as measured by clinical parameters and also to be predictive early on. The early 18 months MRD response was predictive of the outcome at seven years.

So you had that data already in hand before it came to be used as a regulatory endpoint. And I

think the obvious major problem is heterogeneity of the disease in ALL.

DR. REAMAN: I think, as was also discussed, we don't have a good explanation for the mechanism. It clearly is important. But why is it important that the two or three or four drugs, five drugs that you receive in the first 30 days of therapy are so important? And we don't have a good explanation, biologically, mechanistically, for why this is the case.

DR. DEISSEROTH: Yes. Another unanswered question is, what is the cause of the differences one sees in a given therapy, with respect to MRD? Is it the disease? Is it the patient heterogeneity? Is it the therapeutic context, the timing of therapy, or the specific drugs? And I think most people would say there certainly is a combination of the disease parameters --

DR. REAMAN: Plus the host.

DR. DEISSEROTH: -- the host, and the therapy. So one would most likely have to define a very useful therapeutic context in which to test

ALL MRD as a regulatory endpoint.

DR. REAMAN: Elizabeth?

DR. RAETZ: I have a question. So I support fully using MRD as an endpoint for a novel agent, just as you described. But would the converse be true? Because I struggle a little bit. Would you hold each agent to that same benchmark? Because I think there were examples with mitoxantrone in the U.K. trial, where you may not see a beneficial MRD response, but ultimately would see an advantage in EFS.

So could you have it both ways? If you didn't see an MRD benchmark, could you -- or would that be the strict criteria that all agents would be held to?

DR. REAMAN: I think it sort of depends on how the study is designed and the question that's being asked. And from my understanding, with the U.K. experience, looking at mitoxantrone, it was really a randomization between mitoxantrone and idarubicin, and not seeing a concordance between MRD and overall survival.

I think what we're looking for here would really be looking at a single agent compared to a standard, so more of an add-on sort of design, looking at a standard plus a new agent to evaluate, and not considering MRD as a surrogate for comparative efficacy or in the setting of a comparison.

DR. DEISSEROTH: Wouldn't it be important to design and complete the meta-analysis that you were talking about, in terms of defining what is the cut point and what types of therapeutic settings you would be interested in using MRD as an endpoint?

Yes?

DR. GOKBUGET: The meta-analysis which we are planning is actually focused on those patients who have a molecular failure, because we wanted to create a kind of historic comparator group.

We already know that these patients have a very poor outcome, and we don't want to do randomized trials on these patients because we don't have a standard treatment. So the idea is to define a reference group to compare new drugs.

Another remark regarding your point is, why is MRD so relevant, why is the response to these five drugs so important? There is also in vitro data from the Rotterdam group, published in 2009, showing that in vitro response to vincristine, asparaginase, and steroids is highly prognostically relevant.

So ALL is a chemotherapy-sensitive disease. And if it is not sensitive to chemotherapy, the outcome is poor. So for me, it's not too difficult to accept that MRD shows a chemo sensitivity or resistance.

DR. BIONDI: Greg, I think it's very important, what you said before, that the type of design that we're talking is standard plus something versus standard, because that could explain potentially even the different discrepancies in results.

You mentioned the U.K., but even what we heard this morning by Steve, with respect to dexa versus pred and the TKI results, clearly suggest that probably the only way to move on is to prove

the surrogacy of MRD as response activity, to think about this type of design, and to stick on that instead of comparing strategy, which might be, I would say, more complicated, as has been done in the U.K. study, where basically they did compare strategy because they compared two drugs within the same backbone, which I think is a different question.

DR. REAMAN: Dr. Borowitz?

DR. BOROWITZ: So I think the way we're converging is actually very attractive, the idea of having focused studies with a simple addition and not complicating the question, and also looking at an ultimate EFS endpoint along with the MRD surrogate endpoint.

But one of the concerns I still have -- and it's really not exactly this group that has to decide it -- and Steve alluded to this as well. The trial design considerations, particularly in the relapse setting where people go on, if they fail, to all sorts of different things, transplant, different kinds of transplants, different agents,

makes it very difficult to get that ultimate EFS endpoint to be even relatively clean.

Maybe we should rethink our approaches to have a much more comprehensive approach that also deals with not just the relapse -- or the first protocol with the MRD endpoint, but also what happens to patients afterwards, because we tend to lose them. We don't collect data on them, other than maybe whether they relapse or not. And we really need to. If we're going to ultimately solve this problem, we need to actually have access to all those kinds of data when we design the trials. And a lot of the things don't allow us to do that because it's not one trial. It's multiple trials.

DR. VALSECCHI: I was making a consideration on this line, that if we use sort of add-on studies in an accelerated framework for finding new drugs for, let's say, relapsed patients or refractory patients, then we look at MRD as a response. So we look as a measure of activity. And in this sense, it is easier, perhaps, also to foresee a bigger difference than on the final event-free survival.

My question is this. Since, anyway, we would have to ultimately show that also on the event-free survival this has some impact at the end, given the difficulty, then, to standardize -- I mean, the cleanest design would be a real add-on design. So randomized to add a drug to standard or not, and then, however, not to change treatment thereafter until you see the final endpoints of the EFS. Instead, what comes after is often guided by the early MRD response, so transplant is done, yes or no, according to that, and so on.

So what is my question to the regulatory body, and in the experience, may be of other settings -- so a promising drug, let's say, in terms of MRD response, so in terms of early activity, would then be considered for surrogacy with an EFS longer term, let's say observation, even in the presence of different, let's say, interventions after the response, the MRD response.

Is this -- because

DR. REAMAN: There are. And I think Steve

did mention this. And I think, in this setting, the conditions or the requirements, the criteria, for patients moving to transplant may actually be how good is the response to therapy, then the question of surrogacy of MRD in predicting or providing clinical benefit -- if a patient is able to go forward to transplant or not, based on whether they achieved a certain threshold value of a minimal residual disease -- I think is a reasonable question or surrogate.

Michael?

DR. LINK: Yes. But then you have a biased sampling, especially if you're going to make a treatment decision. This is like responders do better. In other words, you have a responder, and then those are the ones that are going to go to transplant. And let's say transplant really makes a difference. I mean, this is still a sore point. But let's say transplant makes a difference. Well, of course, it'll transplant -- let's say just an all or nothing.

Let's say you only transplanted patients who

got an MRD, and let's say transplant really works. Then, obviously, if the drug had any effect on MRD, those patients are going to continue to be in a highly-selected group, and you know your EFS is going to be better, unless you kill everybody with a transplant, which is also possible, by the way.

DR. REAMAN: Which is also possible.

DR. PAZDUR: You do the analysis on all randomized patients from the very beginning.

DR. VALSECCHI: Yes. But if you select -- the problem was if you select the subsequent treatment based on MRD response, then you may have -- well, unless you look then at all the strategy again, not only at the added drug at the beginning, you may have a different --

DR. LINK: Control on transplant.

DR. VALSECCHI: Right, right. The design issues are complicated. That's what I mean.

DR. REAMAN: I mean, those design issues have been with us for decades, and --

DR. VALSECCHI: Yes, we know.

DR. REAMAN: -- irrespective of MRD and

accelerated approval. So I mean, it's been a problem.

I think maybe we have some agreement, consensus -- bad word -- that there is a possibility of evaluating the potential surrogacy of MRD in patients who have failed primary therapy, whether it's clinical relapse, molecular failures, and potentially in high-risk patients. And I guess in this case, I would describe high risk as excluding MRD response to initial therapy.

So patients with sentinel chromosomal translocations that we know predict for a bad outcome would be something for which specific targeted agent development might be appropriate.

But in the setting of newly-diagnosed patients and with the lack of, at least at the present time, a series of randomized questions that are being developed for the induction phase of therapy, we'd be best to really concentrate on the re-induction or high-risk patient population for looking at MRD as a possible response biomarker, qualifying it as such, and then potentially using

it as a surrogate endpoint.

DR. PAZDUR: Can I go to another disease setting here? And that is the very, very, very refractory patients. Frequently, we have sponsors come to us. And because time is of the essence in drug development, they want to develop a drug in a single-arm trial in very, very refractory patients.

They say there's no other therapies for these patients. And there is a small response rate that we see, a small complete response rate. Sometimes, this is then confounded by subsequent, in the single-arms trial, by patients going onto transplant that achieve these CRs. And it's really unclear what is the value of the drug in the single-arm trial.

Does anybody here see any role for MRD in this setting? I'm just bringing that up because that is the reality of our world. Yes. We could talk about these randomized studies, but many people identify a drug, want to get it approved, and they want to get it approved on a single-arm trial in a very refractory disease setting. And

there really isn't a very good way to evaluate this, other than a CR rate.

When we're looking at CR rate, we're really interested also in the duration of those responses. And then they get confounded by introduction of subsequent therapies, transplant, et cetera, even though we're told these patients have no other therapeutic option here, but they eventually go onto these.

DR. REAMAN: I think, here, the very, very, very refractory patients are very, very, very rare. So that's what would really make it difficult.

I think primary induction failures would probably or could be identified as the group we could target. But 98, almost 99 percent of children with ALL -- less adults, but we know why in many cases -- achieve complete remission. So it would be a very difficult study to develop, just based on numbers.

DR. HUNGER: Those trials of which you speak use historical control comparisons, generally. In multiple-relapse leukemia, the historical control

CR rate is X, whatever X may be in their estimation, and want to design a trial where they get more than X.

What we don't have is we don't have good data on the MRD negative -- the percent of patients that attain an MRD negative stayed in the same setting. Now, it's clearly going to be less than the percent of patients that attain a morphologic complete response, probably a lot less.

So if you set a high-enough benchmark -- for instance, if the benchmark for CR rate is 20 percent, as some suggest in that setting -- what if you got 20 percent MRD negative at the end of the first block of therapy? That sounds a lot better than 20 percent with less than 5 percent blasts. So perhaps one could look at it that way.

DR. GOKBUGET: I am familiar with such designs which you've just described because in adult ALL, we actually have such refractory patients. These are, for example, patients with early relapse during treatment, who are refractory to the first line of salvage therapy.

In these patients, the response rates of the first salvage is around 35 percent and to the second salvage, less than 20 percent. And there is no way to compare. There is no comparator arm which you can offer a patient, also due to ethical reasons.

So I really think that the only way in these patients is to do a single-arm trial and to have a historical comparator group. But MRD is maybe an add-on, but the most important endpoint is CR in these patients, in my opinion.

DR. DEISSEROTH: I guess the only idea that I could express here is that the dynamic range of, let's say, very sensitive minimal residual disease techniques could offer you an advantage because with a dichotomous CR or not, you're ignoring what might be happening to various clones within the population. So if you fail to achieve a clinically disease-free state, you're ignoring other events which might be happening beneath the surface.

So it might be useful to think about analyzing MRD in this refractory population to see

if it offered an advantage in terms of identifying active drugs.

DR. GOKBUGET: Absolutely. I will always analyze MRD.

DR. REAMAN: We haven't actually addressed all of these specific questions, at least as they were presented. But I think we've worked through many of the concepts, and I'd like to make sure that we save some time for the discussions on technical issues and concordance between technical platforms, and pros and cons, and standardization, and how we're going to achieve that.

So maybe we could just move to questions 5 through 9 regarding the necessary degree of concordance between flow and PCR methodology in designating critical MRD thresholds.

Should we use a single platform? Is there a role for using both platforms? Is there a role for using one platform and perhaps, in some situations, using another? And maybe we could just have some brief discussion on that issue.

Jacques?

DR. VAN DONGEN: So what platform? I would prefer -- in international studies where we really like to compare results, I would like to propose to go for one platform. And then of course you have to choose the one that is best standardized. That makes everything much more comparable and much more transparent.

That doesn't mean that another platform should not be explored. I'm talking now of course that PCR is more standardized. But I would go, if possible, for flow cytometry. But so far, the standardization level that you reach in PCR is for flow not there. And it still takes an enormous effort.

There is still a lot of subjectivity which should be completely eradicated if you'd like to use it in really multicenter. And by multicenter, I'm not talking about four or five centers. I'm talking about maybe 50 or 80 centers in the U.S. and Europe. If you'd like to do it in that way, it has to be standardized. The subjectivity has to be cancelled. And sensitivity has to be guaranteed,

and the meaning of it has to be the same.

So for PCR, it can be done. For flow, I would say not yet, but I foresee that this should become possible. It needs still a lot of collaboration. So that would be my first reply on this question 5.

DR. LINK: I just have a question. On that presentation, where you saw the discordant results and gave an intermediate group, what was the proportion or percentage of patients that fell into the discordant group category?

In other words, if it's a substantial population, then until you sort of clean up why one works and one doesn't work in that circumstance, you'd almost have to advocate for doing both. If that's just a tiny fraction of the patients, then it's probably less important.

DR. VAN DONGEN: There is not a tiny fraction. There is a shift of at least 20, 25 percent of your patients. That's not a tiny fraction.

DR. LINK: No. I just didn't see the

number. I saw the curves.

DR. VAN DONGEN: No, no, no.

DR. LINK: I didn't see the numbers.

So 25 percent of the patients are discordant one way or the other --

DR. VAN DONGEN: Yes, yes.

DR. LINK: -- more PCR positive and flow negative, both ways.

DR. VAN DONGEN: This is the difference between -- I am not talking about the high risk.

DR. LINK: Yes.

DR. VAN DONGEN: When we would fully focus on the high-risk patients only, then it might be different. On the other hand, the moment that you are going to monitor your high-risk patients and try to get them MRD negative, you run into the same problem, because it's this so-called -- the range, let's say, five times 10 to the minus 4 down to five times 10 to the minus 5, in that range, that is clearly the discrepancy. However, if you also look to the higher levels, it's clear that with PCR, you have another type of definition than with

flow cytometry.

But the most important part, in answering your question, 25 percent is shifting between these low-risk groups and the medium-risk groups, so that's what we have to realize.

DR. LINK: Do we have any such data in doing MRD in the setting like Steve or like has been used in the COG, chaining second remission, the depth of MRD in second remission, based on different kinds of detection?

DR. VAN DONGEN: Not these numbers, not this.

DR. LINK: Because that would really determine how important it is to use both technologies, what's the number of things that are discordant?

DR. VAN DONGEN: Yes, yes. Now, to some extent, you are right, but on the other hand, the moment you have decided to use MRD for quantitation, also after a treatment block, even in a high-risk patient, even in the second remission, then finally, it's about whether you reach a

particular level of MRD positivity.

So it might seem that what is -- then, for the first remission patients, that you cannot directly translate or it's not relevant for the second remission, but finally, it's about the debate. How deep are you measuring your MRD? That's the same principle. That's the same story.

DR. CAZZANIGA: Can I make a comment? Of course it would be preferred to have only one platform because it's easier. But however, we can profit from two different methods applied at different time points to have better results.

To show you, it could be that flow, which is doable at any time point, but the PCR is not doable, for example, at day 15 because there is no sufficient time to do the PCR. So we can use the flow very early to detect not the quality of the remission, but the kinetic of the remission, and then apply a more sensitive method later on for the quality of the remission. So we are measuring different things with different techniques.

This is costly, the logistics are not easy.

And so it's preferable to our eyes in the next future to have one single platform to be used for measuring different things, for example --.

DR. BRAZIEL: Could I make a comment here? I think there is a parallel to what we do in CML follow-up that is exactly what you're talking about. We start with morphology and fish, and standard karyotyping. And then you have a cytogenetic remission, and you go to doing the PCR. You have to go there for the extra sensitivity. So I agree. I think using these techniques together may make some sense.

But I also wondered about the sensitivity of flow going to 8-color or 10-color. Have you looked at that? Because you're talking about 4- or 6-color.

DR. VAN DONGEN: We have indeed looked to whether it increases sensitivity From 6- to 8-color, there is a slight increase in sensitivity. Maybe you can increase half a log. That helps. But whether with 10-color, you gain a half-log, I am not sure.

DR. WOOD: Yes. I agree. I think just increasing the number of colors isn't going to achieve high or large-scale improvements in sensitivity. There may be novel reagents or other approaches that will do that, but I don't think just more colors is the answer there.

DR. HUNGER: I guess I might take the counter-position to my friend Jacques' position, which is that, at least as I understand the concept, we're not talking about optimizing clinical care of patients. We're talking about looking to detect a benefit from administration of a new drug and potential use of that in accelerated, not final, approval.

So to me, you actually don't want to worry about whether you can standardize it between 60 or 80 laboratories, because the fewer the laboratories, the better. I am relatively certain that the agency will be much happier if you come in with an assay performed in a single, or two, or three laboratories -- at least that's our experience --

(Laughter.)

DR. HUNGER: -- than if you come in with an assay performed in 60 to 80 laboratories.

So I think, while the ability to standardize it in many labs might be very useful in clinical management and the big picture of treatment, I think, for an individual clinical trial, confining the assay to as few laboratories as possible is probably beneficial.

I also think, again, you're going to pick some threshold. We could argue what the threshold effect is, but you're going to pick some cut point to call patients positive or negative. And it's not necessarily refining the definition of the intermediate-group patients that's critical in that. It's saying, who clearly didn't respond and who clearly did respond?

So I would actually argue that the design you should look for is as few laboratories as possible. In some cases, perhaps that's flow cytometry because that's the strength of the group. In some cases, perhaps that's molecular technology

because that's the strength of the group. And five years from now, it may be a different molecular technology than it is today, with advents in sequencing technology. But I am not sure we have to worry about standardization in treatment of thousands of patients, because I don't think people are going to go for accelerated approval in trials that involve thousands of patients.

DR. BIONDI: Steve, may I disagree with you in the sense that -- again, I fully agree with what you're talking about. It's not an issue of having 60 or 70 centers. But if you really want to stick to what we said before, on the type of process you want to get as fast-track in introducing and evaluating the efficacy of a new drug, I definitely stand to the consensus of what is the best standardized techniques that we have today, which there is no doubt, at least based on the data, that it is the molecular techniques, which might be adjusted to transient phase; especially if we are talking about a global approval. Because FDA and EMEA are working very close by together, and we

should find a way to speak the same language, at least for a certain time.

Probably today, in terms of methodology, the standardization that has been achieved so far is definitely at least -- maybe I'm wrong, but it's a matter of looking at the data carefully -- in the PCR, although probably more costly, more tedious. But definitely if we are talking about relapsing, refractory ALL study design with standard plus or minus a new drug, assessing response, based on a threshold, and give the number, probably today -- and we want to do that globally because that cannot be done probably in a single content -- I think the only way is to agree upon the same technology, today.

MR. ROBINS: Hi. Chad Robbins from Adaptive Biotechnologies. I just want to make the comment that standardization will be achieved only with a technology that can cut across all disease types or multiple disease types, such as what next-generation sequencing has the promise to be able to offer, if it's one assay that can

characterize and find MRD across multiple disease types. Thanks.

DR. BOROWITZ: So I want to follow up on Steve's thought, because I have to say, until I just heard him say that, I kind of wasn't thinking this way. But I think we should follow this.

In the focus that we're looking at here, which is approval of a drug, remember, we're asking something. We're in a population base. We're going to be able to show a meaningful difference with or without a drug, using a particular assay.

So I guess -- and maybe this is a problem for the statisticians. Maybe it's possible to model how good an assay -- or how bad an assay can be, I mean, to really put a bad face on it -- how bad can an assay be and still allow you to use it to answer the question?

Then we should take the easiest assay that we could do, that meets those criteria, and design the studies. For this purpose, I'm not sure -- I think that if the assay's off 20 percent, but the effects that we're going to need to see to approve

a drug are this big, then maybe that's good enough. I mean, I think we really need to be thinking about these terms for this particular application.

I agree with you, absolutely, Jacques and Andrea, what you're talking about, the discipline of MRD detection and how best to move it out. As a broadly useful tool to assign patients to therapy, and so forth, we need to do all those sorts of things. But for this particular application, how best to approve a drug, we really need to figure out what the performance characteristics of the measurement need to be in order to add it. And it may not need the standardization that we're talking about. Maybe we could use all the assays at different sensitivities, even.

DR. REAMAN: I don't disagree, but I think we'd have to determine the performance characteristics of both techniques and multiple assays, and then somehow determine that they are concordant or can be concordant, and concordant when done in different parts of the world and in different laboratories.

So if that's achievable, feasible --

DR. BOROWITZ: For a population base, they don't have to get the same answer.

I'm sorry. On a whole population base, they don't have to get exactly the same answer on every patient. I mean, Gianni's data show that exactly. They're highly correlated. They both predict. But there are some discordant cases that both predict outcome in both directions.

DR. MARTI: I'm Gerry Marti. I wanted to comment upon a figure that the newest data, I think, we saw this morning, unpublished or submitted, is the 80 percent concordance between the two tests.

This wouldn't be the first time that we have two laboratory tests that are not 100 percent concordant. Discordance is, I think, the more common finding. My thought would be, at whatever time point, 29 days or end of induction, if the patient is flow positive, then you don't have to do anything. If the patient is flow negative, then you have to just consider whether to do the

molecular measurement.

I think someone talked earlier about, historically in CML, before major molecular responses were recorded or became part of the practice, cytogenetics was the key thing. So it's kind of an historical evolution. And it seems like flow is the easier to do and the less expensive, ideally both. Thank you.

DR. REAMAN: But I thought the lack of concordance was in both directions. Right?

DR. CAZZANIGA: Yes. But one way could be to screen with the simplest and less expensive method, and then focus on the negative by a more sensitive method. That could be another way to apply in the two-platform. I would agree on that.

It depends also on the time points, because it could be that it's time-point dependent, because the cutoff is different. If you look at the positivity by flow, it depends whether it's a high or low positivity, because according to the data that I showed, if you have low positivity by flow, it could be also that if those cases are negative

by PCR, as it happened, they are doing better than the cases that have concordant positive.

So it's much more complicated.

MR. CARTER: Chris Carter, Bristol-Myers Squibb. I would ask the committee to consider the intended clinical outcome that would follow from the determination of drug activity.

If you're thinking that the drug add-on therapy is going to lead to cure, that's one thing. If you think that it's going to lead to patients going on to receive stem cell transplantation, for example, you may have a different hurdle that you have to clear in terms of standardization of your assay, going so far as having raised to the level of a companion diagnostic, given the level of risk associated with the subsequent therapy that you're trying to achieve with the new drug that you're adding to the therapy.

DR. REAMAN: I think that's a reasonable point, but I think we're looking to actually define a technique that will give us some confidence across a broad range of potential indications.

I mean, generally, as drugs are introduced in the relapse refractory setting, they find their way into regimens in earlier-stage or more newly-diagnosed patients. So I'm not sure that we can look at how sensitive or insensitive a technique or assay may be for --

I think -- and I certainly would agree that we're not talking about individual patient management decisions. We're looking at a population-based surrogate for efficacy. So maybe there can be some compromise on the degree of sensitivity, but not a great deal.

But I think the issue of how we do this in a multicenter, multinational, multi-continental fashion is important here. So what we may decide to do here at the FDA could and I'm sure will have some implications on what happens with other regulatory agencies.

So we do want to be sure that whatever we do, we can all do it, and that we're all in agreement about how we do it.

Raje?

DR. SRIDHARA: Dr. Sridhara. I think the standardization, I look at in a different way. You do one study where you're comparing a new drug X without the standard therapy. And then you have another study where you're comparing another drug.

If you are using different techniques in these two different studies or if you don't have a standard way of defining what is your response based on this MRD, then in one study, let us say, you see a 20 percent difference. In another study, you see a 30 percent difference.

Are these two the same or are they different? Truly, one is better than the other? You would never know.

So to that extent, you need to say how are you going to define your response. You may be using different technology as to how you're measuring that response, but in the end, it should mean the same thing. Otherwise, in making your clinical decisions, you will be looking at apples and oranges, and you don't know how they compare with each other.

DR. REAMAN: Dr. Becker?

DR. BECKER: So just real briefly, stepping back again from the question of surrogacy, which is what we were trying to deal with in the first few questions and dealing with the technical aspects of it, it might be very difficult to figure out what is that perfect level, that necessary level of concordance between assays.

But there are some things that you can control in the execution of the trial to make sure that you've got a proper balance. For example, you'd want to make sure that arms of patients have equal exposure to both kinds of assays across the different sites at which the assays are being done, that you don't have built-in biases to the way in which patients are being dealt with in the context of the trial, so that unexpected artifacts can creep in, so that even if you can't get to a perfect idea about what's the right analytical validation threshold for comparing one assay to another, at least you can try to control the experiment adequately so that whatever variations

there are between assays can be, to a degree, factored out.

DR. REAMAN: Good point, good point.

I just realized that I missed our scheduled 2:30 break, and I apologize. So we're a little beyond that, so maybe we can take a 10-minute break and resume the rest of these questions and this discussion.

Thanks. So we'll see everybody back at 3:15.

(Whereupon a recess was taken.)

Questions to Workshop Participants, Decisions and Summary (cont.)

DR. REAMAN: We're going to get started with those few of you faithful souls who have stuck around.

So we're going to move on with these questions, but I think, first, Dr. van Dongen wants to talk a little bit about the EuroFlow experience with standardization of flow cytometry. And we'll give you the floor for a short period.

DR. VAN DONGEN: Thank you. The reason I

would like to do this is because of the questions that I heard about standardization and the need for standardization. I really believe we need standardization, whatever arguments.

I cannot believe that the FDA would accept that even if you don't use, let's say, 50 labs across the ocean to do the MRD diagnosis, let's say three or four labs in the U.S. and three or four labs in Europe, that you won't ask for standardization. I guess you would prefer definitely standardization.

I think that PCR now is really for standardized and can almost straightforwardly be used. That's my feeling. I fully agree with all the comments that flow cytometry, for many reasons, has advantages. It's fast. You can use it in the early phase of treatment, et cetera. But there are several main problems, and that is what I would like to address.

EuroFlow went for full standardization on a lot of technical issues, about instrument settings, antibody panels, choice of fluorochromes. I could

give you a lot of examples, and there's a very important other part, and that is software. Because if you go for an enormous amount of data -- I'd like to share this first, to show you about scatter.

This is a scatter pattern about local settings from seven EuroFlow labs before we made any agreement on how to set instrument settings. This is the one with the instrument settings after standardization, which is much more comparable. All these subpopulations are now much more clearly identified.

So then, about handling data, well, there's 3-color analysis or a 4-color analysis you can really think of, oh, this is a precursor B; that's still doable. You can nicely plot your cells against normal patterns. But what is the point if you have 8-color analysis? This is what you cannot read anymore.

First, in EuroFlow, this was a major challenge, to have fast and easy analysis. And fast, I mean really within minutes, and not in

hours. If you see this picture, this is analysis for hours. That should be avoided, and it should be easily standardized.

So we said we should try to work in a completely different way in handling this massive amount of data, and this is what we all know. A massive amount of data from gene expression profiling, we all know. And here you see a typical example from ALL cells in CNS. And this is ALL cells in bone marrow. And we can see that this pattern looks different than the bone marrow cells. So apparently, the ALL cells in CNS are apparently different.

The point is, you can use the heat map, but you also can analyze it in a principal component analysis, like here. This is a principal component analysis, where you use all the parameters in one step. So all these cells, here from the bone marrow, look apparently different than the cells in extramedullary locations like CNS and in testings (ph).

So what we decided at EuroFlow -- and this

is 2006 -- why can we not use this same principal component analysis for flow cytometry? So why, if we now would say this is one leukemia, each dot is one leukemia, if we now would analyze per cell, and then look in the same way as principal component analysis, you get this picture, for instance, in peripheral blood.

So each dot is still a cell, but this is principal component analysis, exactly as here, the same type of methodology. And this pattern is the same if you have standardized your technology in each laboratory. What are we looking at is just T cells, monocytes, B cells, and K cells, analyzed with principal component analysis.

If you do this for one leukemia, there's a comparable 8-color tube. This is one leukemia analyzed from dots and, finally, with the mean value. If you now do this for nine leukemias with the same type with the same tube in principal component analysis, which tries to separate the cells maximally, you see the separation of the same type of leukemias. The same cube analyzed with the

second group of leukemias, you see they nicely split from each other. So nicely, you can use this way of analysis.

You can also use it for quality control, and that's why I'd like to show you this slide. If you now have 30 samples analyzed in the different EuroFlow laboratories, with the same cube, the same instrument settings, you get exactly the same plots, automated. This is just in minutes after acquiring your data. You have exactly the same plots.

So it means that you can really reproduce not subjectively but objectively, as long as you have standardized your analysis. So you discard now all the subjectivity, all the interpretation because the program is doing it for you. And this also means -- and this is the picture I just showed you before -- that you can separate these cells. And this can be done in an automated way, as long as you know where to gate on.

So the only thing that you have to decide is your gating strategy. And that's much less effort

than doing all the removal of all those subjective parameters. And then to get these pictures off, what is the normal differentiation pathway? These classical dot plots, you can transfer in an APS view, fully standardized, and that's what I showed you this morning.

So I believe that this way of analysis, as I showed you, this regenerating bone marrow, can really be done. And it means that flow cytometry can be standardized, and that we potentially can develop 8-color tube approaches, and be able -- I wouldn't say instantly, but in just one or two years' time, have PCR-based approaches replaced by flow cytometry in a fully standardized way, and maybe starting with a few laboratories, but this technology can be transferred to many laboratories. And we have to be aware, some people refer to the CML story, the Bcr-Abl story, that's nice; but be aware, you're talking about transcripts, highly variable. And the IRIS study was performed in a limited number of laboratories. And now, the whole world tries to do exactly the same as these few

highly selected laboratories. And there is a variation that you cannot imagine because it's very, very difficult to perform real-time quantitative PCR on transcripts. It results in an enormous amount of standards around your laboratory. So assuming that CML is standardized, it's not the case.

That is what I would like to say about standardization and the need for standardization. Thank you very much.

DR. REAMAN: Thank you very much.

So, I mean, you've accomplished that within the E.U. Is it something that can be done with flow cytometry in this country?

DR. VAN DONGEN: Yes. I have no doubt. I mean, there is obviously experience here. We are talking about standardization, about tuning things, using the same instrument settings, using the same standard operating procedures.

Then, as I already indicated, what you can easily do, what you see in this picture that's still on, is that you can even have a quality

control round without sending around samples, because if you nicely design -- for instance, for precursor B-ALL, an 8-color tube, or a 10-color tube, whatever, and you use this technical approach and this type of software, then you only take a normal bone marrow sample, where all your precursor B cells are there, and that's your control because all the samples of the different labs should plot over each other precisely. And if something is deviant in your instrument setting, it shifts. Your plots will shift and show that you are not following the right protocol.

So implicitly, you can do a quality control between laboratories by exchanging files, but also by doing exactly the same analysis on a normal sample because the design of all these MRD tubes is including the recognition of normal cells by definition.

So the normal cells are your internal control that you can use between the laboratories regularly, so you can easily say, "Okay. We have quality-control day today. Send in each," I would

say quarter. You send in a sample. The center would collect it. And, yes, your instrument settings and your SOP work is still according to the guidelines. You can check yourself.

DR. REAMAN: Can I just ask you to comment on what you think, and Mike, Julie maybe as well? Is this achievable outside of the European Union?

DR. WOOD: I mean, the idea that one can standardize instrumentation by standardizing instrument settings, and protocols, and the like has been around for a long time. We do this internally in the laboratory all the time. We have three different instruments. They've been standardized in a very similar fashion for more than a decade.

So I don't think that's a new idea, necessarily, but one can certainly, at a technical level, do exactly that. I think that the difficulty has been getting the community as a whole to get together and agree that they're willing to do that. It's not been a technical problem. It's been a political problem.

DR. REAMAN: Politics, we can deal with.

(Laughter.)

DR. REAMAN: So going back to the issue of maybe having both platforms in sort of a step-wise approach, sort of screening, if you will, with flow and then doing PCR, is that a reasonable approach and something that could be accomplished if we don't do one?

Andrea, you say no?

DR. CAZZANIGA: Actually, I would like to get the feedback from the regulatory authority. I would be in a very difficult position to say, "Let's define, both sides of the ocean, an approach to have a selected category of patients, a new method to evaluate the early response and potentially having a fast track," and relying that in different technology.

Again, maybe I'm wrong, but at least, from my perception of the MEA, and not the FDA, that would be very difficult. But maybe I am wrong.

Again, if you go back even to other types that we heard this morning, of a surrogate

biomarker, which has been used in the past for fast approval of some drugs in HIV infections, for example, I would say that probably has been done by using a comparable method.

And so I'm wondering -- well, maybe I'm wrong. One very pragmatic way would be to say let's stick to what is reasonable, standardized today, and in the meantime, because there is no doubt that we need an easier method, work on trying to get a more rapid method, a more affordable one.

I have no doubt. I fully agree with Jacques and Michael that the future will be to rely on this technology, but I have some doubt that the company would agree to do -- at least from my experience. Maybe I'm wrong. Even a recent trial run by a company using the MRD not as a primary endpoint, but as an evaluation criteria of the response, they finally wanted to have a standardized method. And in parallel to that, they do agree to have a second one. Maybe that time will change, but at least today, this is the situation.

DR. WALTON: From the standpoint of

accepting a biomarker as formally qualified, it is the biomarker that we would accept as what is qualified. And as I said earlier, though, that's not independent of the method.

Conceptually, many methods can give us the same result. And if they give us the same result, then I would be happy to have results from any of those assays because I know that had a different assay been run, I'd have gotten the same datum.

If there are assays that give us not quite the same result, then I'd want to know whether or not there is any meaning to the differences between them. So before accepting a mixture of assays, I would want to know, are there differences in the results that one obtains when using the same sample, the two different assays. And if there are differences, do those differences have meaning, or are those differences so small as to be below what has meaning for us? And those would be the questions I would ask.

DR. REAMAN: Thank you.

DR. BOROWITZ: Yes. The only specific point

I want to make is in response to something you just said, Greg, which is, the strategy that was thinking about, of using flow as a screen and if it's negative, to go to a more sensitive assay. I think I'm not in favor of that design, and I think that would be a cumbersome design, not only cumbersome, but a little bit misleading, because the differences between flow and molecular are not just those of sensitivity.

There clearly are discordances within the appropriately sensitive range. And I don't think that that should be a way that we should be looking at it. I think recognizing that broadly these two things must be measuring the same phenomenon, the question is -- I think you've posed the question, and maybe you've answered it perfectly.

What would be acceptable to the administration as an assay that they would approve for a clinical trial? And if the assay has to have the standardized performance characteristics, if that's a primary consideration, then I think Andrea is right. It has to be right now, today. It's

much easier to get an assay approved using PCR because the method is more standardized. As you say, we've seen that in some recent trials. But if that's not the primary -- we really need to know what the agency is going to view as the primary criterion in order to figure out how best to design it. If that's not a primary criteria, then there are advantages to allowing both.

DR. PAZDUR: Let me just jump in here, because there's a bigger issue here. And that is a common principle. Noise obscures the demonstration of efficacy. Sloppiness demonstrates the demonstration of efficacy. Discrepancies between arms and conduct of trials doesn't promote the demonstration of superiority. It obliterates it.

So it's to the company's best interest to have a uniform way of measuring this. Okay? We see this many times. And this is outside of this realm, when we talk about the discrepancies in reading x-rays between an investigator and an independent review committee.

Here again, there are discrepancies of

subjective interpretation of these x-rays, which occurs about, at random, 30 percent of the time. That doesn't help somebody demonstrate efficacy. It makes their life more difficult, from a statistical point of view.

So the companies have every reason -- they're going to be doing superiority trials. They're going to want to have the best chance of demonstrating superiority. And that only occurs with really tight controls of what goes on in their clinical trial. So it's to their best interest in doing this. It's not a regulatory point of view. It's just that demonstration of superiority in a clinical trial.

What we're majorly concerned about is if they demonstrate a demonstration, no matter how they do it, of superiority, is that a true finding? In other words, was a bias introduced here? And that's why, even with the readings of these x-rays, for example, yes, it exists. But if you demonstrate superiority here and win on your trial, is that due to some type of bias. In other words,

was the reading by the investigator at one site or something different than from the IRC, so to speak? In this case, in one arm, the majority of the responses defined by PRC, molecular method, and the other arm by flow.

So those things would make somebody's life more difficult, really.

DR. BECKER: Then the capstone on top of having a well-standardized and unbiased method is, does it reflect, actually, the effect of the drug? Is it a surrogate that does stand in for the clinical endpoint that you care about in the end?

DR. LOZANSKI: I have one global question, because unlike in children ALL, a majority of adult patients with ALL are not treated on clinical trials, and they are treated in community settings.

Therefore, when we are talking here about the development and standardization of MRD as a standard -- perhaps in the future, as a standard of care -- we should consider that a majority of adult patients in the U.S. will be not treated in like supercenters to treat ALL, but in community

settings. Therefore, they probably will be diagnosed and MRD would be probably evaluated in a community setting pathology labs.

Therefore, limiting ourselves to a super-sophisticated molecular method, which requires, as you just mentioned, the big expense and very extensive work-up, and long follow-up before we have results, probably would be not the best option.

I think, if FDA is thinking more globally, it should probably support a study which would help to develop a standardized MRD method based on flow, and not only on a single platform of flow, like BD or Beckman Coulter, but which will be functional for both platforms.

I have also a question for EuroFlow. Was this platform evaluated on a single type of instrument or was it across different instruments?

DR. VAN DONGEN: Also equipment from Beckman Coulter is the dye cosine was tested. And when Beckman Coulter, the new equipment came out, that was late 2009, in fact in the Netherlands, it was

only available in 2010. We had no time to test that equipment again, which we are gladly willing to do so if we get the support of the company, because we have no funds to do testing of equipments.

But all the equipment that was available at the time that we worked on the EuroFlow program were evaluated, so at least four different ones have been tested and they work. There's no problem.

DR. REAMAN: I just remind you again, our focus here isn't on creating what will be an assay that will be used for patient management and promoting what should be or is the standard of care. We're really looking here for what would be the best technique or the technique in which we have the greatest opportunity to eliminate any potential bias for the purposes of approval of a new drug.

I mean, I'm not saying that standardizing tests and regulating tests for patient decisions isn't important. It clearly is, but that's not our

focus here.

So I sense that we are not in sync as to whether we do PCR or flow.

I guess my next question is, how can we best operate -- or is there a way that we can succeed internationally using both platforms and addressing the issue of possible lack of concordance of results between the two?

As Dr. Borowitz asked, is it really identifying a significant problem, based on numbers, degree of sensitivity, when we're not looking at an individual patient's care and management; we're looking at a population-based decision with respect to approval of a drug?

Am I making sense?

DR. BOROWITZ: Can I just follow up on Dr. Pazdur's comment? Because I think that's very revealing. Because I think maybe that really will determine it as what the companies want to do. But I think it's a tradeoff, because the companies are going to have to pay for a clinical trial. They'll either have to pay for a very expensive assay in

that clinical trial or they'll have to pay for more patients to get around the noise problem in that clinical trial. And I don't know, at the end of the day, which one they're going to choose as more worth it from their point of view.

But our only position I think should be to make sure that if they're going to choose among different alternatives, that all the alternatives that they're choosing among are valid.

DR. REAMAN: I'm not sure that just enlarging the clinical trial necessarily addresses the problem of noise. So enlarging the number of patients -- if there's noise, there's noise.

DR. PAZDUR: But I think it is in their benefit, probably.

DR. REAMAN: Right.

DR. PAZDUR: You want to have a uniform platform because here, again, what is going to be the degree of the number of additional patients you might want to enter on here? It becomes an uncertainty. And remember, the major thing in drug development obviously is the reduction of

uncertainty.

So if you can make the uncertainty of the clinical trial, to minimize that, I think, is an important factor in the minds of anybody that's conducting a clinical trial. And to control the things that you can control a priori is something that you want to do. And that's why you have these exhaustive audits that go on in the clinical trial to make sure the B1 on day 21 is collected and the CBC on day 56 is there, et cetera, because they can manage that uncertainty, as far as collecting that data. They can't manage whether a patient has a response or not.

DR. REAMAN: Since we're not going to reach a consensus on the issue of which platform -- and I have to admit, there was some bias in the development of the questions, as I'm sure is apparent to you – maybe we can just talk a little bit about MRD and how analytical performance can be defined, and what's the extent of validation that would be necessary.

So bear with us. We're not eliminating the

PCR platform. But I'd like to just see if we can address some of the issues which remain with a flow cytometry-based assay and how it can be validated, standardized, a performance validation.

So how is analytical performance determined? What panels would be used? We saw several different compositions of antibody panels or number of tubes. How many colors? So can we just talk a bit about that?

Maybe Jacques, you've covered it, but if you and Brent and Michael want to weigh in here, that would be great.

DR. VAN DONGEN: If you talk mainly about flow cytometry, because the question is more related to flow cytometry than to PCR. For PCR, I think the standardization is mainly done. And we have agreed that this could also be performed across the Atlantic. That should not be a problem.

In fact, Julie is joining EuroMRD, so I think that that's problem solved.

If you talk about flow cytometry, in fact, Brent Wood and Michael Borowitz and myself have

already been discussing in other contexts whether this would be possible. And we believe that there are definitely possibilities for technical standardization. That's also what Brent already referred to.

What markers should be useful? Then you ask, in fact, what composition? We agreed that 8 colors would be a logical choice, isn't it?

DR. WOOD: Yes. I think it's a reasonable place to start. It'll certainly encompass the existing assays plus more.

DR. VAN DONGEN: So if 8 colors is a start -- the advantage of 8 colors is that not only a few expert laboratories, but many laboratories on both sides of the ocean can handle it because it's routine equipment nowadays, 8-color machines, laser machines, so that should not be a problem.

The fluorochromes could also be straightforward. But which antibodies? What is the composition? That definitely needs some further investigation. You can make already some logical choices as a start, but this needs further

investigation.

Then my proposal would be, answering everything together, why not start with the PCR-based approach as the standardization in the next period, follow two years. In the meantime, work hard on a fully standardized approach, 8-color approach, so that this is doable at both sides of the ocean in a comparable way; or I would even prefer the words "identical way," so that in two years from now, we speak the same language, have the same message, can actually exchange all the information. That would be the ideal situation, I believe.

But this is hard work. I mean, I know how it is to manage large consortia. I can tell you it's not easy. It's doable, but it's not easy.

DR. CAZZANIGA: May I just ask to Michael, Jacques and Brent the following question? It seems to me that either in U.S. or in Europe, it has been achieved, reasonable and actually very good standardization in flow as well, at least for the data that's been presented this morning.

Then we do have that same in the U.S. and in Europe. And in Europe, we do have this EuroFlow initiative, which to my understanding -- maybe I'm wrong -- is something potentially new that can potentially overcome some of the problems related to the standardization.

So the question would be not to start again, but to address the question how and if this type of approach could be potentially transferrable, because it seems to me -- maybe I am wrong -- that Michael and Brent has a perfect standardization. The flow lab they was presented this morning, Gianni's data on the AIEOP-BFM, Italian and German, worked perfectly well.

But then we do have EuroFlow, which is a third partner. Okay? And the question is, can it be that new partner as a novelty in this field that can potentially overcome? It's a question, maybe a naive question, but I think it's important to get your feedback on that.

DR. WOOD: I mean, at this point, EuroFlow hasn't put forward a proposal for MRD, per se, but

more for initial diagnostic work-up, although they're working on that type of process. We don't have any knowledge of the composition of what EuroFlow is doing.

But I think, in principal, you're right that there's merit in having a similar protocol on both sides of the Atlantic, so we conduct clinical trials. And we just have to agree on what that needs to be. And if it's the EuroFlow protocol, that's fine. If it's another protocol, that's fine. We just need to decide based on the best evidence we have.

DR. VAN DONGEN: I think that the main novelty of EuroFlow is the type of standardization, the control of the standardization, and evaluation during the whole process of running the protocol of the standardization.

So is the implicit quality control completely built in? So the main aim, first, for EuroFlow, was to go for this heavy standardization, talking exactly about the same data, creating the same data, and quality-controlling everything

automatically, and create databases. You could even say databases per protocol, where you have all the data plotted on each other as well.

Then the point that I would say is a very important point available, but that's the second point, what is the composition of your antibodies, that's still in the debate. We tested already a lot. We have a made a lot of progress, but there, we can also learn from each other.

DR. WOOD: Yes. I would just like to say, it would be nice to be involved in the process instead of just being the recipient of other people's work.

(Laughter.)

DR. VAN DONGEN: That's very logical.

DR. REAMAN: Jacques, can I just clarify? The EuroFlow is still based on the premise that this is work that's done in reference laboratories. Right? In specialized laboratories? So this isn't planned to be extended to every institution, hospital, whatever, who may be treating patients on a clinical trial. Correct?

DR. VAN DONGEN: The latter theoretically can be done. If people really have the right level of experience, they can do it. In Europe, there is now a total of 18 laboratories involved, and we will expand to some more laboratories. It is not our intention to make a huge circle of laboratories. We regard ourselves as more developing laboratories, and then by educational workshops to spread the knowledge.

So it could be that there will be much more labs involved. But for a treatment protocol like we are talking about here, for new drugs, I assume that pharmaceutical companies will select a number of labs.

My concern was -- and that's with Bcr-Abl -- as soon as something seems to be standardized, then everybody is going to use this so-called standardized technology, then finally it's not standardized anymore at all. And that's what I would like to prevent.

If you develop something that is excellent for evaluation of novel drugs, fine, but then as

soon as the novel drugs are broadly used, the same technology should be available for all the patients. That's my concern. And that why EuroFlow was initiated.

DR. REAMAN: So we have at least a consensus, I think, on question number 9 -- I'm jumping ahead here -- about the need for centralized performance. And it sounds like, now, that the centralized performance and the standardization of flow, at least with respect to MRD, and the panel of antibodies and some of the other issues is still a process and evolution as it relates to synchrony on both sides of the Atlantic.

So maybe, by default, we are coming to some consensus that if the COG is part of the integrated effort to look at molecular assessment of MRD -- and I don't know if that can be expanded to include the adult groups as well, but I suspect, if it was achievable for COG, it would be achievable for other reference laboratories. We could start there and then proceed with developing the combined approach, Europe, North America, with respect to

flow-based determination of MRD.

Is that a reasonable conclusion?

DR. VAN DONGEN: Yes. Coincidentally, I am chairman of EuroMRD and EuroFlow. As chairman of EuroMRD, I can really say, yes, we are willing also to invite other American teams parallel to Julie Foster's participation.

Of course, Julie knows how much effort it was to do the work, as she visited two centers in Europe, in Berlin and in Monza. But I am sure that this is also feasible for the labs in adult ALL protocols. So if there is a question like that, then I think Julie knows exactly what we are doing and how we are doing things. And we are gladly willing to support that.

For the other step, for flow cytometry, I really would advocate an interaction there because I believe that when we talk also the same language in flow cytometry, it would be a major step forward and particularly to have this also done across the ocean. I would really be in favor for that. And if there is the question there, I will support it.

DR. REAMAN: Mike?

DR. BOROWITZ: So I think one critical piece that's a little bit different about the flow and the molecular harmonization, when you think about EuroFlow, is, in the case of PCR, you're really standardizing what's exactly the same assay.

The proposal in EuroFlow is to standardize what's exactly the same assay, but right now, both AIEOP-BFM and the COG have validated functional assays that have worked on thousands of patients, where we have outcome data.

So I think, really, the first step, rather than just saying we're going to go to a standardized flow assay, is to really do some validation of a proposed new assay against our existing assays to investigate their performance characteristics before saying that the place to go is to a new assay, which is focused on precision, which it does very well, but we don't have the hard data that says that it's accurate in terms of predicting outcome.

That will come with time, but I'm not sure

we need -- it may be more efficient to do some work standardizing the two assays together before we say that that's going to be the gold standard, around which we're going to circle.

DR. VAN DONGEN: Yes. I would like to comment on this. You're talking about assays and suggesting now that there might be three assays. And I believe there are not three assays. What EuroFlow is doing is much more focusing on the standardization procedure and, of course, the combination of the 8-color tubes, that's an important part. But that's something you can do together, that you can maybe choose together.

But what I'm talking about is the integrated approach of full standardization of instrument settings, as a piece in your laboratory, and the quality control in such a way that this is very straightforward and easy. So that's not an assay. That's a complete strategy.

DR. BOROWITZ: Right. No. I'm not questioning that part. But EuroFlow also has a software component that uses novel parameters to

analytically identify what you're calling MRD. That's completely novel and not validated in any -- it may be far superior for reasons you put together, but it's certainly not validated in any sense, in a way that's similar to what those of us who have been doing it by the old method have already done. And I don't think that's, as they say, ready for primetime as an assay without really doing -- the precision aspect of it and the standardization aspect, that's perfect. But that's not sufficient to make that the assay of record, where it stands now.

Adjournment

DR. REAMAN: Then I think that's why deciding to maybe start with the molecular approach, a PCR-based approach, and taking some time to develop whatever the gold standard turns out to be -- we might have a platinum standard that's even better -- I think makes some sense, at least as an interim way to proceed, if that's what everyone's in agreement with.

Well, I'm sensitive to the fact that people

have international flights that they have to connect or get to and that there are cabs coming. And I'm not going to do a formal summary, but I think there are some important points on which we've reached consensus.

Some issues which remain, principles that can be used as action items for continued work, investigation discussion, and hopefully collaboration. But I think this has been a great discussion, and I appreciate all of your efforts to be here, to participate, and especially those of you who have traveled great distances, even having to overnight it in Minneapolis. I consider that a great distance.

(Laughter.)

DR. REAMAN: So thanks, everybody.

(Whereupon, at 4:04 p.m., the meeting was adjourned.)