THE LEARNING HEALTH SYSTEM SERIES

CARING FOR THE INDIVIDUAL PATIENT

Understanding Heterogeneous Treatment Effects

David Kent, Jessica Paulus, Mahnoor Ahmed, and **Danielle Whicher,** *Editors*

NATIONAL ACADEMY OF MEDICINE

THE LEARNING HEALTH SYSTEM SERIES

CARING FOR THE INDIVIDUAL PATIENT

Understanding Heterogeneous Treatment Effects

David Kent, Jessica Paulus, Mahnoor Ahmed,
and **Danielle Whicher,** *Editors*

NATIONAL ACADEMY OF MEDICINE

WASHINGTON, DC
NAM.EDU

NATIONAL ACADEMY OF MEDICINE 500 Fifth Street, NW Washington, DC 20001

Support for this publication was provided by the Patient-Centered Outcomes Research Institute® (PCORI®), through two awards: a Patient-Centered Outcomes Research Institute (PCORI) Eugene Washington PCORI Engagement Award (1900-TMC) and the Predictive Analytics Resource Center (SA.Tufts.PARC.OCSCO.2018.01.25); and the Predictive Analytics and Comparative Effectiveness (PACE) Center at the Tufts Medical Center. The views presented in this publication are solely the responsibility of the editors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors, or its Methodology Committee; or of the PACE Center and/or the Tufts Medical Center.

International Standard Book Number-13: 978-1-947103-16-0
Library of Congress Control Number: 2019948398

Printed in the United States of America

Suggested citation: Kent, D., J. Paulus, M. Ahmed, and D. Whicher, Editors. 2019. *Caring for the Individual Patient: Understanding Heterogeneous Treatment Effects.* Washington, DC: National Academy of Medicine.

ABOUT THE NATIONAL ACADEMY OF MEDICINE

The **National Academy of Medicine** is one of three Academies constituting the National Academies of Sciences, Engineering, and Medicine (the National Academies). The National Academies provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on issues of health, health care, and biomedical science and technology. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

Learn more about the National Academy of Medicine at NAM.edu.

WORKSHOP PLANNING COMMITTEE

DAVID M. KENT (*Chair*), Tufts Medical Center
THOMAS CONCANNON, RAND Corporation
ROBERT GOLUB, *Journal of the American Medical Association*
SHELDON GREENFIELD, University of California, Irvine
RODNEY HAYWARD, University of Michigan
A. CECILE J. W. JANSSENS, Emory University Rollins School of Public Health
MUIN J. KHOURY, Centers for Disease Control and Prevention
PETER ROTHWELL, University of Oxford
EWOUT STEYERBERG, Leiden University Medical Center
ANDREW J. VICKERS, Memorial Sloan Kettering Cancer Center

NAM Staff

Development of this publication was facilitated by contributions of the following NAM staff, under the guidance of J. Michael McGinnis, Executive Officer and Executive Director of the Leadership Consortium for a Value & Science-Driven Health System:

DANIELLE WHICHER, Senior Program Officer
MAHNOOR AHMED, Research Associate
JESSICA BROWN, Executive Assistant to the Executive Officer
JENNA OGILVIE, Communications Officer

Tufts University Staff

DAVID KENT, Tufts Medical Center
JESSICA PAULUS, Tufts Medical Center

Consultant

ROBERT POOL, Hired Pens LLC

REVIEWERS

This Special Publication was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with review procedures established by the National Academy of Medicine (NAM). These reviewers were asked to consider the accuracy of the content within this Special Publication, the accuracy with which conversations at the workshop on which this Special Publication was based were conveyed, and the strength and balance of this Special Publication's arguments.

We wish to thank the following individuals for their contributions:

FRANK DAVIDOFF, *Annals of Internal Medicine* (Emeritus)
SETH MORGAN, National Multiple Sclerosis Society
JODI SEGAL, Johns Hopkins University
CHRISTINE STAKE, Ann & Robert H. Lurie Children's Hospital of Chicago

The reviewer composition includes individuals with subject-matter expertise, attendees at the workshop, and those who did not attend the workshop. The reviewers listed above provided many constructive comments and suggestions, but they were not asked to endorse the content of the publication, and did not see the final draft before it was published. Review of this publication was overseen by **DANIELLE WHICHER,** Senior Program Officer, NAM, and **J. MICHAEL McGINNIS,** Executive Officer, NAM. Responsibility for the final content of this publication rests entirely with the editors and the NAM.

PREFACE

"The premise of traditional research is to put a treatment at the center of consideration and decide, Is this treatment helpful for an average patient? Trouble is, there aren't very many average patients out there, and I, like most people, am not an average patient."

—Seth Morgan, neurologist, multiple sclerosis patient, and patient advocate

Evidence-based medicine (EBM) arose from a clear need and represents a major advance in the science of clinical decision making. Traditional approaches to decision making based on expert opinion, extrapolations of pathophysiologic reasoning, or personal experience led to extreme variations in practice patterns, which have been well documented, starting in the 1970s (Wennberg and Gittelsohn, 1973). Many routinely accepted clinical practices have been found to be ineffective (or harmful) when subjected to evaluation by randomized trial designs, and large proportions of "effective" procedures were found to be inappropriate when scrutinized by expert review (Chassin et al., 1987). More broadly, it is well understood—not only in medicine, but in many fields—that human decision making is plagued by fundamental cognitive biases, and that statistically driven decision making has general advantages compared with human "expert" judgment (Kahneman et al., 1982; Meehl, 2013).

Despite broad acceptance of EBM, however, a fundamental incongruity remains unresolved: Evidence is derived from groups of people, yet medical decisions are made by and for individuals. Randomization—introduced by R. A. Fischer in the field of agriculture and ported into clinical research by Austin Bradford Hill—ensures the comparability of treatment groups within a clinical trial, which allows for unbiased estimation of average treatment effects. If, like farmers growing crops, we treated groups of patients instead of individuals, or if patients with the same disease were identical to one another in all factors that determined the harms and the benefits of therapy, then these group-level averages would make a perfectly sound foundation for medical decision making. However,

patients differ from one another in many ways that determine the likelihood of an outcome, both with and without a treatment. Nevertheless, despite persistent assertions by clinicians that determining the best therapy for each patient is a more complicated endeavor than simply picking the best treatment on average, popular approaches to EBM have encouraged an over-reliance on the average effects estimated from clinical trials as guides to decision making for individuals.

Shortly after the turn of the 20th century, the decoding of the human genome promised to deliver us from one-size-fits-all medicine. But a decade and a half later, it appears unlikely that genetic information will be leveraged broadly or deeply into clinical decision making. The effects of individual single nucleotide polymorphisms (SNPs) tend to be small (Goldstein, 2009), they typically add little information to easily obtainable clinical or phenotypic information (Ioannidis, 2009), and even in combination they account for only a small proportion of heritability (Manolio et al., 2009). (The limitations of polygenic scores are well reviewed in A. Cecile J.W. Janssens's presentation; see Chapter 4.) While more than 350 different pharmacogenomic associations are included in pharmaceutical labels, the clinical utility of these tests is generally not established; and despite important efforts (e.g., those described by Josh Peterson; see Chapter 5), pharmacogenomics has not brought us substantially closer to understanding individualized benefit–harm trade-offs for most interventions.

Notwithstanding the challenges of unraveling the genetics of disease states and the disappointments, to date, of gene-based approaches to diagnosis, prognosis, and treatment, the goals of personalized medicine remain deeply compelling. Better population-based outcomes will only be realized when we understand more completely how to treat patients as the unique individuals they are. Our patients surely expect nothing less. The reality of effect modification (i.e., that the same treatment in different patients may have different consequences) is undeniable to any physician. For example, angiotensin inhibitors can both cause and prevent kidney dysfunction, anticoagulation treatments can both cause and prevent strokes (hemorrhagic and embolic, respectively), and antihypertensive medications can both cause and prevent cardiac events. But these patient-level variations are not completely unpredictable. A simple medical history and physical examination can provide abundant information about how patients with the same disease (or those included in the same trial) can differ from one another in many important ways that influence benefit–harm trade-offs.

In May 2018, under the auspices of the National Academy of Medicine (NAM), we gathered a group of experts and stakeholders—physicians, methodologists, patients, payers, and regulators, among others—to discuss the tension between group evidence and decision making for individuals. The group focused on

"predictive" approaches to heterogeneous treatment effects (HTE). That is, for evidence to be more applicable at the individual patient level, we need to combine methods for strong causal inference (e.g., randomization) with methods for prediction that permit inferences about *which* particular patients are likely to benefit and which are not.

One point of agreement for better patient-centered evidence was that rather than serially examining subgroups defined "one variable at a time" for statistically significant interaction effects, a more relevant approach is to disaggregate patients by fundamental dimensions of risk using models that incorporate the effects of multiple prognostically important clinical variables simultaneously to yield "personalized" estimates of benefit–harm trade-offs. Risk dimensions that are important for decision making include the risk of the primary outcome of interest (as patients at higher risk often have greater potential for benefit) and the risk of treatment-related harm. Disaggregating patients into strata defined by these risks can yield information about effects that may be obscured in the overall average and in conventional subgroup analysis. Another important point of agreement was that information on both harms and benefits of treatment across these different risk strata should be presented on an absolute scale—rather than a relative risk scale—to support clinical decision making.

While the principles for these "predictive" HTE analyses of randomized controlled trials were introduced more than a decade ago (Kent et al., 2010; Rothwell et al., 2005), speakers at the conference noted that recent developments and refinements in such analyses provide reasons for optimism, including the investment of more resources in patient-centered outcomes research (particularly through the Patient-Centered Outcomes Research Institute [PCORI]); the priority PCORI has given to research accounting for HTE; advances in "big data" in medicine (and in the broader culture) that facilitate development, validation, and continual updating of prediction models; new methods for prediction using machine learning (discussed by Fan Li; see Chapter 4); new adaptive research designs developed to cope with and leverage patient heterogeneity (discussed by Derek Angus; see Chapter 2); the broad dissemination of electronic health records (EHRs) and incentives for their "meaningful use"; specific support in the Patient Protection and Affordable Care Act for shared decision making; and the "open data" movement encouraging new models for clinical trial data sharing, enabling individual patient meta-analysis capable of supporting well-powered predictive HTE analysis.

Additional dimensions of evidence individualization discussed herein include the need for effective implementation strategies for the use of prediction models that promote physician and patient acceptance (discussed by John Spertus; see

Chapter 5); developing new quality measures to incentivize personalized care that transcends binary all-or-none rules, which tend to promote low-value care (discussed by Rod Hayward; see Chapter 5); enhancement of restrictive formularies to permit doctors and patients the latitude to select pharmaceuticals that work best at the individual level; and new value frameworks for pharmaceutical pricing that take this heterogeneity into account (discussed by Robert Dubois; see Chapter 3).

Despite substantial progress and many points of agreement, the workshop also highlighted numerous controversies, challenges, and research gaps. These included determining the appropriate role for observational data, understanding the comparative performance of machine learning methods compared with traditional statistical approaches for predicting HTE, and developing guidance on methods for assessing the effectiveness or validity of models that predict *benefit* (i.e., the *difference* among potential outcomes with alternative treatments, rather than just predicting outcome and prognosis).

In summary, there was broad agreement that while the challenges remain formidable, a better understanding of the heterogeneity in treatment effects has the potential to truly transform medical care, improve health outcomes, and reduce unnecessary or ineffective therapies by targeting treatments to those most likely to benefit. The discussions captured in this volume are critically important for moving this conversation—and medicine in general—forward in the decades to come.

We would like to thank all of the attendees at the workshop on which this Special Publication is based for their generous and robust conversations. We would also like to thank Mahnoor Ahmed and Danielle Whicher of the NAM, Jessica Paulus of Tufts University, and Robert Pool of Hired Pens LLC, all of whom, along with David Kent, contributed significantly to the drafting and editing of this Special Publication.

David Kent, M.D., M.S.
Director
Predictive Analytics and Comparative Effectiveness (PACE) Center
Tufts Medical Center

Joseph Selby, M.D., M.P.H.
Executive Director
Patient-Centered Outcomes Research Institute (PCORI)

J. Michael McGinnis, M.D., M.P.P.
Executive Officer
National Academy of Medicine

CONTENTS

BOX AND FIGURES

BOX

FIGURES

SUMMARY

Medicine is currently undergoing a paradigm shift from evidence-based practice to a personalized approach. A shortcoming of evidence-based medicine (EBM) is that it lacks precision by applying broad-based group data to the treatment of an individual. Yet, each patient is unique, and treatment responses differ from one person to the next. This variability in treatment response is called heterogeneous treatment effects (HTE), the study of which is essential for doctors to effectively tailor treatments for their patients to maximize the benefits while minimizing the harm.

On May 31, 2018, the National Academy of Medicine, in conjunction with the Predictive Analytics and Comparative Effectiveness (PACE) Center at the Tufts Medical Center, held a workshop in Washington, DC, to discuss approaches to examining HTE to personalize and improve patient care. Funded by the Patient-Centered Outcomes Research Institute (PCORI), the day-long discussion centered on the following motivating questions:

- **Potential:** How can clinical trial data be analyzed to yield reliable patient-centered treatment effect estimates? What are the state-of-the-science methods for assessing treatment heterogeneity?
- **Risks:** How can we be sure personalizing evidence will improve decision making, as compared with the default of relying on overall average treatment effects? What are the evidentiary standards for implementing changes to clinical practice to personalize care based on evidence of HTE?
- **Lessons learned:** What can be learned from the challenges of genomics-based personalized medicine? What can be learned from the efforts of previous clinical trialists to understand more personalized treatment effect estimates?

- **Strategies:** How should clinical research and clinical practice be redesigned to support the generation and the dissemination of patient-centered evidence?

This publication summarizes the remarks and the insights of workshop participants consisting of patients and patient advocates, physicians, medical researchers, research funders, and health insurers, as well as representatives from pharmaceutical companies, federal agencies, professional associations, and medical journals. The conversation began with a discussion of the promise of exploiting HTE to personalize care for patients, the related key concepts and considerations, examples of the types of analyses that have been conducted, and challenges for the field. One challenge with modeling treatment effects is identifying an appropriate reference class or group of patients with a similar set of characteristics to reflect the target patient. For many reasons, it is now recognized that conventional subgroup analyses that examine how treatment effects vary across characteristics "one variable at a time" are of extremely limited value for informing care decisions. Defining subgroups based on outcome risk has emerged as a useful (if imperfect) approach to separating the patients most likely to benefit from a treatment from those unlikely to benefit or those most likely to experience net harm. Ultimately, the goal is to develop sophisticated composite risk scores that reflect a range of patients' personal variables such as comorbidities, functional status, mental health status, and the various social determinants of health.

As researchers and clinicians search for ways to best deal with and take advantage of HTE, they must consider patients' needs and preferences. For instance, patients need to understand the relationship between the average treatment effect described in clinical trials and their own individual situation. Additionally, it would benefit patients if studies of a specific condition were conducted in a uniform way that enabled results to be compared across studies—and for trials to be pooled to provide the statistical power needed to describe variations in treatment effect. Given these priorities, a push toward patient-centered care will undoubtedly alter the traditional relationship between patients and health care systems, with patients playing a more active role in their care.

The transformation is not exclusive to patients. Regulatory agencies and health insurance companies have to rethink their assessment of medical treatments. The U.S. Food and Drug Administration (FDA) looks for a variety of differences in how people respond to drugs based on variables such as demographic differences, genomic characteristics, and disease severity. For payers, the assessment of medical treatments and the examination of treatment heterogeneity have reimbursement implications, especially in the current environment in which payers are relying

more frequently on value frameworks to determine which treatments to cover for which cohort of people.

To deal effectively with HTE, there are several methods and models for predicting how individuals will respond to different treatment approaches. Using multiple genes to predict predisposition to a disease, polygenic risk scores have been used since the late 1990s to identify high-risk groups for targeted interventions. In recent years, however, the predictive performance of polygenic risk scores has come under question, with numerous studies proving their inability to clearly distinguish which groups of people will likely develop a disease from those who will not. Machine learning offers an additional set of analytical tools. With advancements in computing power, machine learning methods (e.g., penalized regression, regression tree-based methods, Bayesian nonparametric models, ensemble learners) make it possible to spot correlations in data that are beyond human capacity. Yet, despite the theoretical appeal of these methods, applications of these tools in general practice have been limited.

This is just one of the many barriers to implementing HTE prediction models and techniques in routine clinical care. Apart from ensuring clinical validity, HTE predictive models need to demonstrate clinical utility and workflow advantages. Prediction tools must be able to integrate seamlessly into a medical records system so as to provide clinicians with near-real-time results and improve decision making. Addressing these issues is necessary to impress provider confidence in these tools. Without physician acceptance, HTE models will be meaningless and will fall short of their potential to improve the value of care.

As highlighted by participants' remarks, the field of HTE is still in its infancy. It must not only address outstanding methodological questions, but also determine best practices for implementing risk models and predictions tools in clinical practice. Therefore, key directions for the field include

- Developing guidance on approaches for assessing the effectiveness or the validity of predictive and prognostic models;
- Understanding the comparative performance of supervised machine learning methods that can be applied to understand HTE;
- Facilitating collaboration and leadership across various sectors of the research ecosystem to create prioritized opportunities for large trial re-analyses or collaborative individual patient data analyses to examine HTE most likely to impact population health;
- Describing approaches to implementing risk models in clinical care and providing guidance on which approaches are most effective at informing decisions both at the point of care and at the level of the health care system;

- Considering approaches for integrating data related to the social determinants of health into risk-prediction models;
- Determining the role for observational data and when it is appropriate to combine randomized controlled trials and observational data;
- Reforming the predominant fee-for-service payment system in the United States to one that rewards value and population health improvements;
- Promoting dissemination of innovative trial designs, including those sampling larger and broader populations to enrich patient heterogeneity; and
- Establishing or extending research reporting guidelines to promote the conduct of predictive HTE analyses.

Addressing these priorities will require deliberate coordination among a wide range of stakeholders, including researchers, clinicians, payers, regulators, health delivery organizations, and medical journals, with the ultimate goal of serving the patient. The individuality of the patient should be at the core of every treatment decision. One-size-fits-all approaches to treating medical conditions are inadequate; instead, treatments should be tailored to individuals based on heterogeneity of clinical characteristics and their personal preferences.

1

INTRODUCTION

"I will go out on a limb and predict that this is the most important meeting you will attend this year, at the National Academy of Medicine or elsewhere." So said Joseph Selby, Executive Director of the Patient-Centered Outcomes Research Institute (PCORI), in his opening remarks at the National Academy of Medicine (NAM) workshop titled Evidence and the Individual Patient: Understanding Heterogeneous Treatment Effects for Patient-Centered Care. While Selby's opening comments were intentionally provocative, they indeed captured the clear sense of many attendees that the workshop topic was both timely and extremely important.

Heterogeneous treatment effects (HTE) refer to the way effects of a treatment can differ, sometimes dramatically, from one patient to the next. While such variation, or heterogeneity, can be quite challenging to clinicians, who would find their jobs easier if every patient responded to a treatment uniformly and predictably, heterogeneity also offers great opportunities. The challenge, then, is to learn how to transform those opportunities into concrete benefits for patients.

In his introduction, Selby discussed the relationship between HTE and the field of evidence-based medicine (EBM), which seeks to firmly ground medical practice in the strongest possible evidence, such as data from randomized controlled trials. One problem with EBM, Selby said, is that doctors are often expected to apply evidence-based recommendations to all patients. If, for example, randomized controlled trials show that the average patient with high blood pressure will benefit by lowering blood pressure to below 120/80 mm Hg, then doctors are expected to work to get the blood pressure of all their patients under that level.

Further acknowledging this problem, he said,

> There has been a nagging sense that we weren't quite getting it right. And there has been a huge backlash from physicians who in the 1980s and 1990s encountered evidence-based medicine for the first time and said, "But you're not

any longer allowing me to do what is basically my job, which is to personalize the treatment for the patient in front of me and to consider, particularly, their risks."

The study of HTE, however, offers the potential for doctors to once again personalize treatments for their patients.

OVERVIEW OF THE WORKSHOP

The workshop, held on May 31, 2018, at the National Academy of Sciences building in Washington, DC, convened physicians, medical researchers, representatives from funding agencies, health insurance companies, pharmaceutical companies, federal agencies, professional associations, and medical journals; as well as patients and patient advocates, to discuss approaches to leveraging health data to examine HTE in order to personalize and improve patient care (see Appendix C for the complete workshop agenda). By understanding the reasons for the heterogeneity and developing ways to predict how individual patients will respond to a treatment, medical researchers and physicians should be able to personalize medicine to a far greater degree than is possible today. Such an ability would open the door to treatments that are more effective with fewer side effects and would also allow patients to make more informed decisions about the types of medical treatments they choose to receive.

That is the potential of understanding HTE, as many workshop participants commented. But to reach that potential will require advances on both the research side and the clinical side. To explore those requirements, the NAM, in conjunction with the Predictive Analytics and Comparative Effectiveness (PACE) Center at the Tufts Medical Center, convened this workshop, with funding from two awards from PCORI.

Participants were asked to consider four motivating topics over the course of the meeting (see Appendix B for list of workshop participants, web participants, and staff):

- **Potential:** How can clinical trial data be analyzed to yield reliable patient-centered treatment effect estimates? What are the state-of-the-science methods for assessing treatment heterogeneity?
- **Risks:** How can we be sure personalizing evidence will improve decision making, as compared with the default of relying on overall average treatment effects? What are the evidentiary standards for implementing changes to clinical practice to personalize care based on evidence of HTE?

- **Lessons learned:** What can be learned from the challenges of genomics-based personalized medicine? What can be learned from the efforts of previous clinical trialists to understand more personalized treatment effect estimates?
- **Strategies:** How should clinical research and clinical practice be redesigned to support the generation and the dissemination of patient-centered evidence?

In addition to these questions, there was an explicit recognition that there are additional questions of central importance to patients. Not only will patient cooperation be critical in the design and performance of clinical trials that aim to understand HTE, the patients themselves will also inevitably be partners with clinicians in making treatment decisions about their care whenever HTE are present. As several workshop participants noted, it will be important for patients to have a clear understanding of HTE in order to make informed choices about their care. With this in mind, the workshop participants were asked to think about HTE from the point of view of the patient and to consider the following questions that patients might ask:

- Given my personal characteristics, conditions, and preferences, what should I expect will happen to me?
- How can I use knowledge about HTE to improve the outcomes that are most important to me?
- How can clinicians, as well as the care delivery systems they work in, help me make the best decisions about my health and health care?

The day-long workshop was divided into five sessions, each with individual presenters and responders, as well as a discussion session that followed the presentations. This NAM Special Publication provides a summary and synthesis of the presentations and the discussions that took place during the workshop. Its structure mirrors that of the workshop, with each of the five chapters reflecting one session.

Chapter 2 provides an overview of HTE, introducing concepts, examples of types of analyses that have been done, and illustrations of how their application has led to more individualized clinical decisions. Chapter 3 summarizes a discussion with patients, patient representatives, and other stakeholders regarding the importance of understanding HTE. Chapter 4 examines methods that can be used to produce models that will predict treatment effects, with a discussion of the strengths and the weaknesses of the various approaches. Chapter 5 delves into the issues involved with implementing clinical programs that take HTE into

account. The final chapter, Chapter 6, offers a look to the future, addressing what will be required to account for HTE in medical practice.

The opinions expressed by workshop attendees and reproduced within this publication are those of the individual speakers and are not the position of the National Academies of Sciences, Engineering, and Medicine or the NAM. Workshop presenters and participants were not asked to come to any consensus opinions, and any recommendations made were those of individuals, not the group as a whole. However, there were various areas in which there was apparent widespread agreement among those at the workshop, and those areas are noted, as appropriate.

2

THE PROMISE OF PERSONALIZED EVIDENCE-BASED MEDICINE

After the introductory remarks, the workshop began with a session that explored the potential to take advantage of heterogeneous treatment effects (HTE) to improve and personalize patient care. Several presenters described how understanding this heterogeneity can lead to more effective treatments for individual patients, thus maximizing benefits and minimizing harms. Much of the information offered in this session was relevant to the patient question: Given my personal characteristics, conditions, and preferences, what should I expect will happen to me?

Points Highlighted by Individual Speakers

- Patients in randomized trials typically vary substantially in their risk of the primary study outcome. Because of this, patients also vary in their harm–benefit trade-offs. The average results from a clinical trial may not even reflect the trade-offs of the majority of the patients in the trial. (Kent)
- Risk-based analysis can help separate out the patients most likely to benefit from a treatment from those unlikely to benefit or those more likely to experience net harm. (Kent)
- It is valuable to develop decision scores that score patients on expected net benefit, meaning expected benefit minus expected harm. Examining a clinical trial with the aid of such a tool can provide insight into the outcomes of individual patients in the trial. (Basu)
- To understand heterogeneous treatment effects, it is important to get past the one-variable-at-a-time analysis and take into account the fact that there are usually multiple axes of heterogeneity. (Angus)

USING RISK-BASED FORECASTING TO PERSONALIZE MEDICINE

"What we're really talking about is personalized evidence-based medicine," said David Kent, Director of the Predictive Analytics and Comparative Effectiveness (PACE) Center at the Tufts Medical Center. In other words, the goal is to use evidence from randomized controlled trials (RCTs) and other sources to predict what is likely to happen with an *individual* patient. He and the other panel members discussed two types of prediction: outcome risk modeling, that is, creating models that differentiate patients by risk, and treatment effect modeling, or separating patients by the likely effects of treatment.

Doctors and medical researchers have long recognized the limitations of RCTs for providing evidence for clinical decision making. Indeed, Kent said, even Austin Bradford Hill, who pioneered the use of RCTs in medical research, commented 50 years ago that while RCTs can determine the better treatment on average, they "do not answer the practicing doctor's question, What is the most likely outcome when this particular drug is given to a particular patient?"

The innovation of evidence-based medicine (EBM), Kent continued, was the realization that RCTs could be used by doctors to determine what is best for individual patients, which required what he called a "very subtle" shift in approach. Instead of seeing RCTs as tools for establishing causation, they were now seen as tools for prediction in single cases. But single-case prediction is a problematic area, he said, and "a lot of very smart people have thought deeply about it." Kent mentioned in particular Nobel Memorial Prize in Economic Sciences winner Daniel Kahneman, who identified two distinct approaches to such a prediction. One is the "inside view," which looks at the specifics of a case, weighs the various factors, and then synthesizes them into a prediction. "This is the view that physicians had before evidence-based medicine," Kent said, and it "is really the view that we spontaneously adopt for making decisions in virtually all aspects of life." The second approach is the "outside view." In this case, predictions are made by explicitly identifying a group of patients with similar diagnoses and characteristics, known as a reference class, and using that reference class as a statistical basis for prediction.

In contrast to traditional medicine, EBM relies on the outside view. Specifically, EBM is a type of reference class forecasting. "It relies on making inferences for single cases based on the frequency of outcomes or estimated treatment effects in a reference class to which the individual of interest is similar," Kent explained. Yet, this raises another problematic question: How does one define similarity? He referred to this situation as the classic "reference class problem," which was first

described in 1876 by the mathematician John Venn, who noted that each item or event has a multitude of attributes that could be used as the basis for categorizing it into one class or another. How do you choose from that multitude? For doctors, making that choice is a real problem, because determining the class to which a patient belongs will have implications for his or her treatment choices.

"How does evidence-based medicine approach this very deep problem, the reference class problem?" Kent asked. "Generally, I think we've largely ignored it. What we've done is we've emphasized the broadest possible reference class, which is the overall effect in a trial." On the other hand, one can quickly run into problems when dividing patients into groups according to various characteristics. "If you have just 10 binary attributes, then you have over 1,000 unique subgroups that you can describe," he said, "and if you have 20 attributes, you have over 1 million subgroups that you can describe. And you quickly run into the problem of small sample sizes."

What is needed, he explained, is a principled way to prioritize which attributes are important in determining both the outcome of interest and the benefits of therapy. He and his colleagues have suggested that one particularly useful approach is to define subgroups according to outcome risk. Regardless of how treatment effects are measured (i.e., as the absolute risk reduction or as a relative risk reduction), the control event rate is a mathematical determinant of treatment effect—and the control event rate is simply an observable proxy of the outcome risk. When the outcome risk varies substantially across different groups of patients in a trial, the benefit–harm trade-offs are also likely to vary substantially.

To explain why outcome risk is a valuable way to classify patients, Kent presented a figure displaying absolute mortality risk as a function of a patient's percentile mortality risk for patients with acute myocardial infarction (see Figure 2-1). Specifically, he said, the figure depicts patients with an ST-segment elevation myocardial infarction, a type of heart attack caused when the coronary artery, which has been affected by atherosclerosis, is blocked by a blood clot at the site of an injury. "This hockey stick–shaped distribution is actually a scatter plot with 1,000 little dots, each representing a patient," Kent explained.

As shown in Figure 2-1, the risk of death averaged over all medically treated patients is 6 percent, which, according to Kent, is the number that would appear in a typical analysis. "The control event rate would be 6 percent," he stated. That percentage, however, obscures some critical details. For example, the fact that, when risk is determined by a multivariable model using easily obtainable baseline clinical variables, 75 percent of patients actually have a risk that is lower than the average, and 50 percent of patients have a risk that is only half of the average rate—that is, the median patient has a mortality risk of only 3 percent. Furthermore,

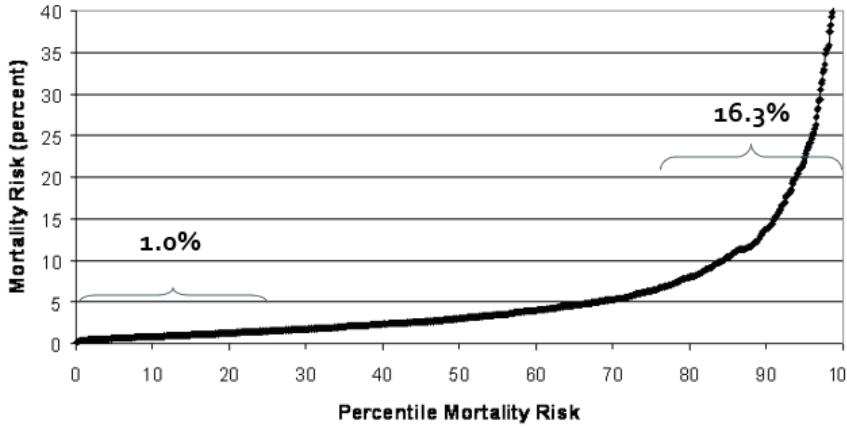


FIGURE 2-1 | Distribution of mortality risk in medically treated patients with acute myocardial infarction.
SOURCES: David Kent presentation on May 31, 2018; Kent et al., 2002.

at the extremes, the differences among patients are pronounced. The lowest-risk quartile of patients has an average mortality risk of only 1 percent, while the highest-risk quartile has an average mortality risk of 16 percent (see Figure 2-1). "Doctors actually know that the risk–benefit trade-offs in these patients are different," Kent noted, "but in the trial, they're all lumped together."

To further illustrate the value of stratifying patients by risk, Kent presented an analysis of how two risk-stratified subgroups fared in the Danish multicenter randomized study of fibrinolytic therapy versus primary angioplasty in acute myocardial infarction, known as the DANAMI-2 trial (see Figure 2-2). DANAMI-2 analyzed 1,572 patients who presented to a hospital with an ST-segment elevation myocardial infarction, or STEMI. Some patients were treated with pharmaceuticals to break up the clot, while others were treated with percutaneous coronary intervention (PCI), in which a catheter is used to insert a stent and open up a clogged artery. Figure 2-2 shows the long-term results of PCI versus clot-busting drugs in two groups of patients studied in DANAMI-2: the lowest-risk quartile and the highest-risk quartile from the distribution in Figure 2-1. "The high-risk patients, the minority of patients who are high risk, get tremendous benefit from PCI compared to medical therapy," Kent explained. "But the majority of patients who are low risk [are] actually slightly harmed by PCI compared to medical therapy." If you combine results from all groups, the benefit to high-risk patients overwhelms the harm to low-risk patients, and PCI appears to always be the superior choice.

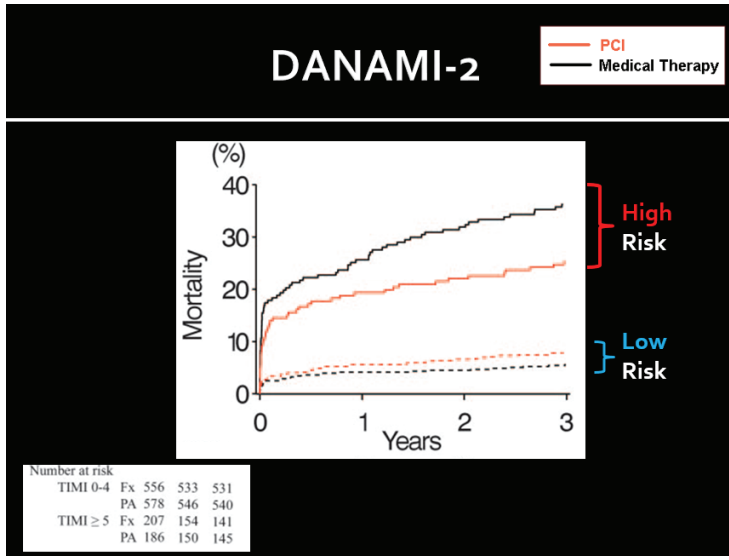


FIGURE 2-2 | Results of percutaneous coronary intervention (PCI) versus medical therapy (tPA) in DAMANI-2 for high- and low-risk patients.
SOURCES: David Kent presentation on May 31, 2018; Thune et al., 2005.

The researchers who published the results of DANAMI-2 analyzed one-variable-at-a-time subgroups (e.g., groups defined by age, sex, race, or the presence or absence of diabetes or hypertension) and found that the same overall benefit existed. "Just like every other trial," Kent said, "they claimed consistency of effects, but that's because they didn't stratify by risk." On the other hand, contrasting groups of patients who differ by only a single variable under-represents the heterogeneity found among patients. As in many trials, he said, if you separate your analysis into high- and low-risk subgroups using multiple risk factors, you may observe results similar to those that appeared in the DANAMI-2 trial.

Kent then described how he and his colleagues analyzed 18 randomized treatment comparisons by studying the effects on patients separated into quartiles according to risk. When they examined the trials on the basis of relative risk (i.e., risk in the treatment group divided by risk in the control group), there were no clear patterns. But when they analyzed the trials on the basis of absolute risk, fairly consistent patterns emerged, with those in the higher-risk groups receiving greater benefit from the treatments. And, indeed, the analyses of three of those trials were deemed clinically important enough to be published in three separate clinical papers (Kozminski et al., 2015; Sussman et al., 2015; Upshaw et al., 2018).

In one of those papers, published in *The British Medical Journal*, Kent and his colleagues analyzed the results of the Diabetes Prevention Program (DPP) RCT

(Sussman et al., 2015). In that trial, 3,060 nondiabetic patients with evidence of impaired glucose metabolism were randomized to one of three groups: a group that was given metformin, one that was given a lifestyle intervention, and another that received usual care. The main outcome measure was whether a patient developed diabetes. Kent and colleagues showed the risk-stratified results calculated in two ways. The first was as a hazard ratio, for which the risks of a treatment group are compared with the risks of the control group to determine a measure similar to relative risk. When examined by the hazard ratio, the effects of the lifestyle treatment were homogeneous—people in every risk quartile benefited by about the same amount—approximately a 50 percent relative risk reduction. By contrast, the effects of the metformin treatment group were heterogeneous. The lowest-risk group saw no benefit whatsoever, while the highest-risk quartile obtained about a 50 percent relative risk reduction, and the intermediate quartiles received something in between.

"We have one intervention where the statisticians will say [there is] no heterogeneity of treatment effect, and another where there is," Kent summarized. Notably, when the DPP results are shown on an absolute risk difference scale versus a relative risk difference scale, which is clinically the most important measure of treatment effect, there are important HTE for both interventions. These results further demonstrate the "scale-dependence" of HTE; whether it is present or absent depends on what scale is used to describe treatment effects. "And for both interventions," Kent noted, "it may be important to make different decisions for different patients and to target the treatments to the high-risk groups, particularly if resources are in some way limited."

In one final example, Kent discussed a re-analysis of the Digitalis Investigation Group (DIG) study (Kozminski et al., 2015). The DIG study was an older trial in which more than 7,000 patients with heart failure were given either digoxin or a placebo, with the outcome measures being hospitalization due to heart failure and all hospitalizations. Patients in the highest-risk quartile experienced nearly a 15 percent absolute decrease in hospitalization due to heart failure when given the digoxin versus the placebo, while those in the lowest-risk quartile experienced only a 2 percent decrease. "But when you throw in all hospitalizations," Kent said, "you see something interesting." He further explained, "If you look at the lowest-risk quartile, you see that there's actually harm. And this makes sense because digoxin has a very low therapeutic index, and these are patients who really can't benefit because they're not at risk for hospitalization. They can't benefit, but they can only get the toxicity that sometimes causes hospitalization with digoxin. So, there's actually net harm in those patients." Once again, if these results were only analyzed in the conventional way, this important heterogeneity in benefit–harm

trade-offs would be obscured both by the overall results and within conventional (i.e., one-variable-at-a-time) subgroup analyses.

In summarizing, Kent offered the following take-away messages:

- Overall benefits-to-harm results may be driven by a relatively small group of influential (typically high-risk) patients.
- The typical (median) risk patient is frequently at a considerably lower risk than the overall average.
- The average benefit seen in the summary results often over-estimates the benefit (on the absolute risk difference scale) in most patients and may obscure harm in many others.
- Risk-based subgrouping is often clinically informative and usually feasible.

Finally, he noted several caveats and a few thoughts on how to proceed:

- Outcome risk is not the ideal subgrouping variable.
- Ideally, researchers would model outcome risk with therapy *versus* without therapy, incorporating all important treatment effect interactions; but modeling treatment effect interactions has its own challenges.
- Risk-based subgroup analysis can avoid these problems because it is performed blinded to treatment assignment.
- Researchers and clinicians can either use an external, already-developed model to stratify patient populations by outcome risk or they can develop an endogenous (or internal) model blinded to the treatment effect; either of which may avoid much of the troublesome issues associated with more aggressive data-driven approaches.

DEVELOPMENT OF A DECISION SCORE TO OPTIMIZE TREATMENT DECISIONS

In the next presentation, Sanjay Basu, Assistant Professor of Medicine at Stanford University, spoke about a variation on the risk-based analysis that Kent described. In particular, Basu and his colleagues created a decision score that considered a patient's expected benefit from treatment, as well as the expected risk, in order to assess the expected net benefit from treatment. Their analysis made it possible to make sense of two major studies of blood pressure treatments that had arrived at different conclusions and to predict which patients would do best with which treatment approaches.

The original question arose, Basu explained, because of two studies that appeared in *The New England Journal of Medicine* 5 years apart. The first,

published in 2010, reported the results of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial (ACCORD Study Group et al., 2010), which had a total of 4,733 participants who were followed for nearly 5 years. The study looked at the value of using intensive blood pressure control to keep people's systolic blood pressure below 120 mm Hg, as opposed to the standard goal of keeping blood pressure below 140 mm Hg. Basu stated that the study concluded that targeting a systolic blood pressure of less than 120 mm Hg "did not reduce the rate of a composite outcome of fatal and nonfatal major cardiovascular events." The patients in the study arm with intensive blood pressure treatment did not improve more, on average, than the control patients who had the standard treatment.

Five years later, in 2015, results were reported for the Systolic Blood Pressure Intervention Trial (SPRINT; SPRINT Research Group et al., 2015), which also examined the value of using aggressive blood pressure treatment with a target systolic blood pressure of 120 mg Hg versus a standard treatment with a target of 140 mg Hg. The conclusion of SPRINT, however, was diametrically opposed to that of ACCORD. Basu stated, "[T]argeting a systolic blood pressure of less than 120 mm Hg, as compared with less than 140 mg Hg, resulted in lower rates of fatal and nonfatal major cardiovascular events."

There was one obvious difference between the trials. "The first trial was among people with type 2 diabetes, and the second was not," Basu noted. Nonetheless, evidence from several other trials indicated that the presence of diabetes did not have a profound enough effect to explain the two trials arriving at such radically different answers—which left clinicians in a bind. Which trial should they trust? Various editorial writers offered differing opinions. Perhaps there were differences in the sample selection between the two trials. Perhaps the presence of type 2 diabetes had a larger effect than previous studies indicated. Or perhaps, Basu said, "HTE exist, and despite being part of an overall similar population, differences in sampling resulted in a somewhat different average treatment effects between the trials."

This discrepancy was not only an academic issue. In particular, SPRINT found the intensive-treatment group was significantly more likely to suffer severe adverse effects (e.g., hypotension, syncope, electrolyte abnormalities, acute kidney injury or failure) than those in the standard-treatment group. "This is not such a benign choice for the primary care physician," he stated. "Rather than simply being a matter of causing some nausea or headaches, the side effects of intensive treatment may in some cases be severe: hospitalization, disability, dialysis, and death. So, one would want to make the right decision even though blood pressure control may seem like a fairly benign treatment decision," Basu further explained. With this in

mind, Basu and colleagues decided to analyze the two trials in attempt to explain the discrepancy in average treatment effect between them in terms of HTE. "Perhaps," he said, "similar patients in predictable ways have more benefit than harm, and vice versa, and differences in sampling could lead to differences in the average."

The research question guiding their study was, Which patients have the most potential for benefit and the least potential for harm from the intensive blood pressure intervention? Their analytical approach to answering that question involved developing two Cox regression models, one for benefit (i.e., a reduced risk of cardiovascular events and deaths) and one for harm (i.e., an increased risk of severe adverse events). They also chose a limited set of potential candidate variables based on previous studies that indicated potential reasonable factors that might influence the HTE. Among these candidate variables were demographic characteristics, tobacco use, pre-randomization laboratory values, medication use, and systolic and diastolic blood pressure. The model also included a term for treatment and treatment by covariate interactions. In an effort to reduce false positives, Basu and his colleagues used an elastic net regularization approach with repeat cross-validation with subsamples of the data. Collinearity was also found to be a problem, as many of the variables were interrelated. With many collinear variables, Basu said, the solution is either to choose one variable that can stand in for all of them or to shrink the coefficients among the many collinear variables.

In an a priori specification, they decided to separate people in terms of their net benefit, which was equal to the benefit of the intensive treatment minus the harm. They then created a benefit–harm score based on clinically accessible variables such as age, sex, race, systolic blood pressure, number of blood pressure medications taken, use of aspirin or statins, tobacco use, serum creatinine, urine microalbumin and creatinine, and total cholesterol and high-density lipoprotein. Next, they applied the benefit–harm score to the participants in SPRINT, retroactively assigning them "decision scores" for the trial, and then compared those decision scores with the real outcomes of the trial. What they found was that the SPRINT participants with the higher decision scores were more likely to have benefited from the intensive treatment and less likely to have experienced harm than those participants with lower decision scores.

When Basu and his colleagues divided the SPRINT participants into tertiles based on the decision scores, they observed distinctly different patterns of response to treatment between the top and the bottom tertiles (see Figure 2-3). In the top tertile—that is, the one-third of those subjects whose decision scores indicated they were most likely to benefit from intensive treatment—participants who received intensive treatment had much greater benefit than the control subjects who received the standard treatment. There was no difference, however,

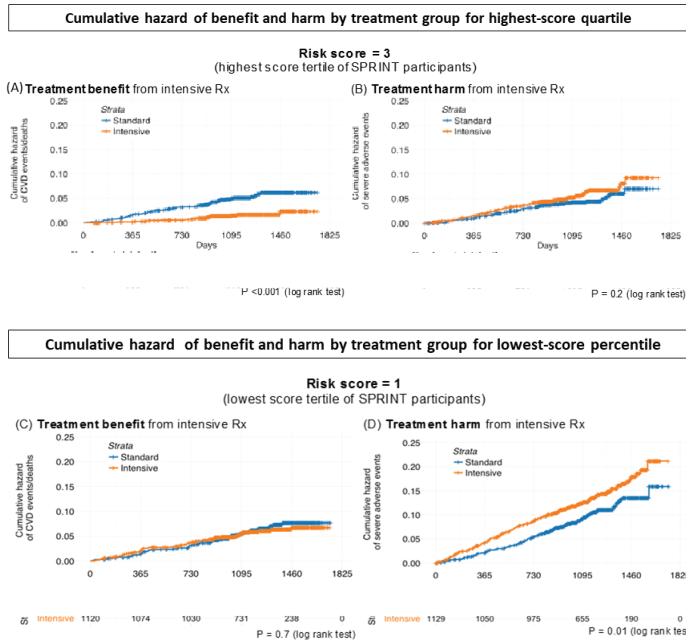


FIGURE 2-3 | Treatment benefit and treatment harm for highest (a) and lowest (b) tertile of SPRINT participants, based on their net benefit decision score results.
SOURCES: Sanjay Basu presentation on May 31, 2018; Basu et al., 2017.

between the standard group and the intensive group in the amount of harm they experienced in the form of adverse effects. Thus, among the highest tertile there was a significant net benefit to treatment.

Conversely, among the lowest tertile—that is, those whose decision scores indicated they were least likely to benefit from aggressive treatment—there was no difference between the intensive-treatment group and the standard-treatment group in the benefit they received from the treatment in terms of reduced cardiovascular events and deaths. But among those in the lowest tertile, subjects in the intensive-treatment group experienced significantly more harm (i.e., adverse events) than those in the standard-treatment group.

Next, Basu and his colleagues applied the decision scores to the ACCORD subjects and found the same pattern. Among the highest tertile on the decision score, the intensive treatment had a net benefit versus the standard treatment; but among the lowest tertile, the intensive treatment had a net harm (these data are not shown in Figure 2-3). What explained the different results from the SPRINT and the ACCORD studies?

Although the average effect for SPRINT was positive (i.e., intensive treatment led to better outcomes on average) and the average effect for ACCORD was neutral or negative (i.e., intensive treatment did not have better outcomes on average), the outcomes of both trials, when examined more closely, were in fact not that different. The perceived variance was due to the difference between the samples for the two studies, in terms of their likelihood for net benefit from aggressive blood pressure lowering. First, as Basu explained, although 21 percent of the ACCORD sample was predicted—and observed—to benefit from the aggressive therapy, there was a larger percentage of SPRINT subjects who fell into this high-benefit group. In the end, the decision score derived from the SPRINT study correctly predicted that most ACCORD patients would not benefit.

The true lesson from the two trials, Basu concluded, is that "average trial results can often hide clinically profound heterogeneities in treatment effects." Average trial results may also appear to be contradictory, consequently confusing both clinicians and the public. Comparing the average effects of SPRINT and ACCORD overlooked vital details about how individuals can be expected to respond to blood pressure treatment; in particular, the aggressive blood pressure treatment could be expected to help only a subset of patients—not all of them. Furthermore, Basu said, it was necessary to consider several factors in combination, rather than any single factor, in order to explain the important variations.

Several limitations of the study were noted by Basu. Their analysis could not examine results further than approximately 5 years, because SPRINT was discontinued after that amount of time. Another limitation was that congestive heart failure could not be included as a negative outcome because of differences in definitions between the two studies. People may also weigh benefits and harms differently in their calculation of net benefits. Basu and his colleagues are now exploring other approaches to weighting benefit and harm, rather than simply treating them equally.

Basu believes it will eventually be possible to create a tool that makes treatment recommendations for individual patients based on their individual characteristics and preferences. As an example of how such a tool would assist doctors, he noted the difficulty in keeping track of which of the numerous available drugs for treating type 2 diabetes are best for which type of patient—who might either benefit or be harmed by each type of drug. He said,

> What we're experimenting with in a trial setting is doing a personalized risk estimate for a baseline risk. Does the patient want to be treated or not, or how aggressively might we think about doing treatment? That's the classical absolute risk before treatment. And then from individual participant data and network

> meta-analyses, we can calculate heterogeneous treatment effects across all the possible treatments that are available, [identify] what types of people might benefit more or less from each different type of therapy, and then weight it based on patient preferences.

People are different, he noted. There are some patients, for example, who simply will not inject a medication; others are willing to inject a drug, but may be worried about weight gain or avoiding hypoglycemia. The ultimate goal is to use these various factors as weights to create an individualized ranking of medications based on the individual's personalized risk and preferences and, particularly, the uncertainty in those estimates. "That, I think, is on the horizon," he concluded.

DESIGNING RANDOMIZED CONTROLLED TRIALS WITH HETEROGENEOUS TREATMENT EFFECTS IN MIND

Derek Angus, Chief of the Department of Critical Care Medicine at the University of Pittsburgh, opened his presentation with the image of being on a Scottish mountaintop, where it is possible to look around and see everything clearly in all directions. "And that is ideally where we want to get" in HTE, he said. "We want to have some sense of the exact therapy that the patient would absolutely want and be most likely to benefit from." However, he said, it is not so easy.

"We look out over this cloud inversion, and every valley around us is filled with clouds, and as soon as we walk off the top of the mountain, we end up in a very unique valley filled with clouds, and everyone tries to solve the problem for navigating inside just that valley and comes up with a solution that appears to be partly solving the problem—but not all of it." That is the current situation with HTE. Everyone is grappling with just part of the problem.

Actually, it is quite difficult to combine the HTE approach with the precision medicine approach and the patient-centered approach. To provide some context, he quoted from a paper by Richard Kravitz and colleagues that examined the role of HTE in EBM (Kravitz et al., 2004). The authors, Angus related, identified four dimensions of HTE:

- Baseline probability of incurring disease-related event;
- Responsiveness to the treatment;
- Vulnerability to adverse events; and
- Utilities (expressed by patients, maybe society) for different outcomes.

Historically, most HTE papers have focused on the first and third of these dimensions, Angus said. "As they go down into their valley, they make some assumptions." He noted that Kent stated in the previous presentation, in essence, "We'd like to predict response to treatment, but we're going to just predict risk of having the disease-related event." Conversely, those interested in precision have tended to concentrate on the second dimension. "It comes from people who feel they understand the disease on the inside, and so they've tended to focus on response to treatment," he explained. Furthermore, there is a whole field whose researchers focus on the utility of different outcomes. Each group tends to work in its own separate valley.

For the duration of his presentation, Angus discussed the relevance of the design of RCTs for studying HTE. As Kent previously described, HTE analyses seek to identify various subgroups who respond differently to treatment, with the higher-risk subgroups having larger absolute treatment effects than the lower-risk ones. The typical risk distribution in these clinical trials is left-shifted, as the majority of the participants fall at the low end of the risk axis (see Figure 2-4). In these cases, Angus noted, the median risk is always lower than the average risk.

A major challenge in analyzing such trials is not overlooking those low-risk subjects who, in addition to not receiving any benefit from the intervention, are actually harmed. A typical one-variable-at-a-time subgroup analysis, as Kent also noted, will generally miss this harm. Comparing treatment effects in men versus women or Caucasian versus African American subjects will uncover a relatively small range in net benefit. "Therefore," Angus said, "you want to have

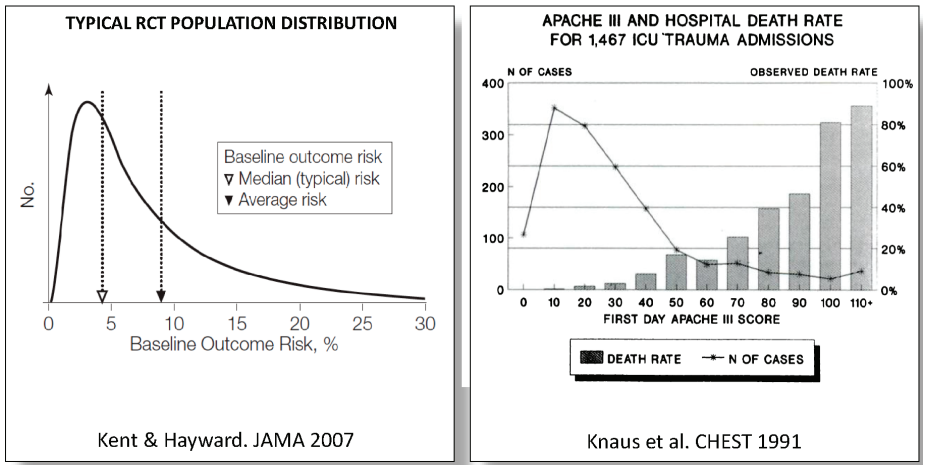


FIGURE 2-4 | Typical risk distributions in clinical trials are left-shifted.
SOURCES: Derek Angus presentation on May 31, 2018; Kent and Hayward, 2007; Knaus et al., 1991.

this multivariate risk model that spans across the entire range, where you can have quantiles far to the left" that will identify subgroups of subjects who are harmed by treatment. "Of course, this will require having huge sample sizes all the time, enrolling across the entire breadth of the disease of interest, so that we always have enough samples to build these models," Angus explained. "And so, the answer to trialists is just to do huge trials—enrolling everyone at risk."

With regard to precision medicine, he described how researchers in that field tend to think more in terms of prognostic and predictive biomarkers. A prognostic biomarker is one that provides information about the likelihood of a patient reaching a certain disease-related endpoint, while a predictive biomarker is one that offers information about the likelihood of a patient responding to a particular therapy. Both biomarkers provide useful information for personalized medicine; that is, for a treatment to be useful for a particular patient, that patient must, first, be likely to experience the effects of the disease and, second, be likely to respond to the treatment.

A single biomarker is not necessary, Angus said. Indeed, it is possible to use a suite of biomarkers to identify patients most likely to respond well to a particular treatment. As an example, he described a study in which the researchers used principle component analysis on a large quantity of biomarkers to define two phenotypes (Calfee et al., 2014). The study was similar to a multivariable analysis in that the researchers analyzed a large number of biomarkers; but the end result was assigning patients to one of two categories, just as in a one-at-a-time variable analysis. "They were very happy with themselves," Angus noted, "because these phenotypes were obviously not predicted clinically," and still the phenotypes were useful both prognostically and predictively. "If you have phenotype 2, you were much more likely to die. At the same time, the same phenotype was highly predictive of benefit versus harm when exposed to the different strategies."

"This is the essence of much of the precision medicine world—trying to get at these predictive biomarkers," he said. "But they just seem to have forgotten this lesson learned from HTE about the peril of having a subgrouping based on a single variable" because of the way it may hide issues with people who are in the lowest-risk quantiles.

Yet another problem arises, Angus said, in the way that unmeasured baseline variables can cause huge differences in patient outcomes. The net effect of a treatment can jump from harm to benefit or vice versa with modest swings in the prevalence of these unmeasured variables.

"The problem here is that you think you're studying one disease, but you're not really," he said. "So what can be done?"

People in the precision medicine field are approaching the problem in a couple of ways, he said. "They basically have what I would call the 'hope and pray' models. If they think there's a complex disease, they may have some putative biomarkers, and they either ignore the biomarkers or they take a bet ahead of time and only enroll on the biomarker." He said he would not speak of those further.

Instead, he turned to what he called the "spread the bet" models. "You acknowledge you do not know everything about the intervention and you also do not know everything about the disease, and you're going to try to learn as you go."

The best and most evolved version of this approach, he said, is the adaptive platform trial. Such a trial focuses on a disease, not a particular treatment; it uses multiple interventions (in multiple arms) with continuing enrollment; it is often based on Bayes' theorem, a formula that describes how to update one's hypothesis as new information is uncovered; and it involves tailoring one's choices over time. So far, he continued, researchers using adaptive platform trials have been focused on the pre-approval space in drug testing, and the emphasis has been on efficiency, with the trials relying on small sample sizes. Different therapies "graduate" to the next phase while the trial continues.

The "poster child" for adaptive platform trials, Angus said, is the I-SPY 2 trial that screened several promising breast cancer therapies simultaneously. The first results came out about 18 months ago, with papers published in *The New England Journal of Medicine* (Carey and Winer, 2016; Park et al., 2016; Rugo et al., 2016). Patients are assigned to different arms of the trial with "response-adaptive randomization," which regularly changes the selection rules according to the results of the trial to that point. As an example, Angus described how a planned 400-person trial might proceed. If, after results were available for 40 patients it was clear that treatment A was looking much better than treatment B, the randomized selection rules would be modified so that a greater percentage of the next 40 subjects would be put on treatment A (see Figure 2-5). "You don't have to be an investigator" to make that call, he explained. "It can be a preset algorithm."

An advantage of this approach is that if indeed treatment A is superior, it will become statistically clear sooner, and the study can be stopped earlier than planned. On the other hand, if the apparent advantage of treatment A in the first 40 patients was because of random chance, then the next 40 patients will move the outcomes back toward 50/50, and the trial will continue. One caveat, Angus said, is that this is not very efficient for a two-arm trial because the power is still determined by the smaller group. "But it actually becomes very interesting in the situation where you have multiple arms and multiple subgroups, which is arguably the situation we're in today [with heterogeneity of treatment effects]."

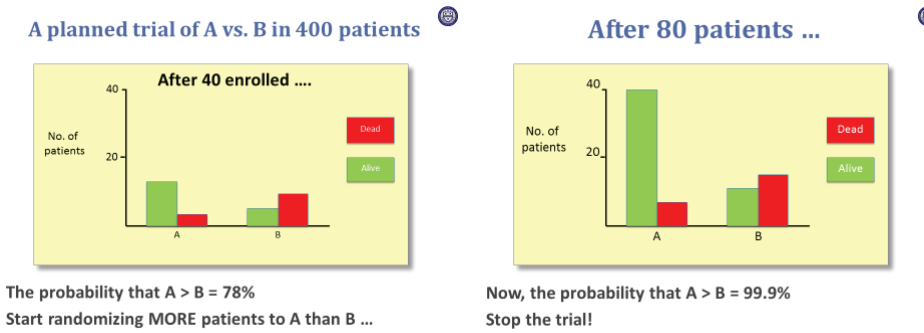


FIGURE 2-5 | In adaptive platform trials, a promising treatment can be more quickly validated.
SOURCE: Derek Angus presentation on May 31, 2018.

The use of the adaptive platform approach was successful in the I-SPY 2 trial, Angus said. When the trial began, there was uncertainty about which drugs worked and in whom they worked. What they found was that the use of one drug, neratinib, was effective only in patients with one of two different combinations of the three biomarkers used in the trial; while a second drug combination, veliparib–carboplatin, worked only in women with a different combination of the three biomarkers (Park et al., 2016).

In conclusion, Angus reiterated several points. First, there are generally multiple axes of heterogeneity, and one of the challenges is to keep this in mind and not restrict the problem down to a single axis. The classic HTE literature has largely focused on the baseline risk of disease balanced against a constant threat of avoiding one-variable-at-a-time subgroups in favor of multivariable risk models. Precision medicine studies largely ignore that one-variable-at-a-time warning and instead concentrate on "predictive" biomarkers that may not actually predict. They use trial designs with putative predictive enrichment and "hope and pray" that it works. The alternative "spread the bet" approach is quite exciting—it is working in cancer and is arguably more patient-centered.

REGULATORY UTILITY OF UNDERSTANDING HETEROGENEOUS TREATMENT EFFECTS

Robert Temple, Deputy Center Director for Clinical Science, Center for Drug Evaluation and Research at the U.S. Food and Drug Administration (FDA), highlighted the importance of understanding HTE from the regulatory perspective. He offered several additional examples in which drugs had contrasting effects in

different patients and in different situations, further emphasizing the critical need to understand treatment heterogeneity in order to maximize the benefit of drugs.

He began by commenting on how the field's increasing knowledge of the pharmacokinetics of drugs has led to a better understanding of why various subgroups of patients may respond differently to the same medication. "Forty years ago," he said, "you didn't know how a drug was metabolized, you didn't have good evidence of how it was renally excreted, hepatically modified; we didn't understand about the enzymes that were responsible for the drug's metabolism." To illustrate, he mentioned the case of the tricylic antidepressants. Tricyclics must generally be given at a dose of 150 to 300 milligrams (mg) to work. Yet, years ago people did not start on that dose—they started with 30 mg—because some people had terrible adverse effects on 150 to 300 mg. Why did such reactions occur? Some individuals are simply poor 2D6 metabolizers (i.e., their CYP2D6 enzymes do not function well) and they do not metabolize the tricyclics as quickly as most people. Consequently, these people will have approximately five times as much of the drug in their bloodstream as a "normal metabolizer" given the same dose. "If you just gave them the 300 mg, you could kill a poor metabolizer because those drugs are toxic at high doses," Temple said. "So, the standard starting dose for desipramine was 30 mg, the right dose for a poor metabolizer. If that worked okay, you were fine. If it didn't but was tolerated, you increased the dose. Of course, delaying effective antidepressant treatment poses its own problems." Fortunately, this scenario is no longer an issue, he noted. "We know most of the metabolizing enzymes, and we know how to adjust doses for people. In clinical trials, we get blood levels on almost every patient, so we can detect unanticipated reasons for some people to have higher blood levels than others."

According to Temple, FDA now looks for a variety of differences in how people respond to drugs—"differences in how you metabolize the drug, differences because of a concomitant drug that affects the metabolism of a drug, or differences in pharmacodynamic effect, that is, differences in how some people respond to the same blood level, perhaps because of genomic differences." Today, when FDA receives a drug application, it examines all possibilities that might affect either safety or effectiveness, including demographic differences, genomic characteristics, and severity of the disease. "And, every once in a while, those analyses of subgroups turn up something important," he said. "I'll give you two of my favorite examples."

His first example was BiDil, a combination of isosorbide and hydralazine that is used for heart failure. BiDil was examined in two studies by the U.S. Department of Veterans Affairs (VA), and the overall results showed that it performed a little

better than a placebo and much worse than an angiotensin-converting enzyme (ACE) inhibitor (Cohn et al., 1986, 1991). However, when they looked at the drug's effectiveness broken down by various demographic groups, the researchers observed a surprising result. The drug did not perform well in Caucasian subjects, but it was effective in African American subjects (Carson et al., 1999). "That was true in both studies," Temple said. "And we eventually allowed a confirmatory study to be done entirely in [an African American] population, and the effect size was very dramatic." This is one of many discoveries that can result from analyzing subgroups, he noted.

The second example concerned ticagrelor, an alternative to clopidogrel, which is an antiplatelet drug that was used in people who had experienced a heart attack. A large cardiovascular outcome study revealed that the drug worked better than clopidogrel everywhere except in the United States, where it was considerably worse (Wallentin et al., 2009). "When we examined the data," Temple said, "it turned out that it was entirely attributable to the dose of aspirin that was used. When it was used with 300 mg of aspirin, it performed worse than clopidogrel, but when it was used with 100 mg of aspirin, it performed markedly better." And aspirin use was distinctively different in the United States, where about half of the patients used the 300 mg dose; in contrast to the rest of the world, where only about 15 percent were given that dose. Thus, differences in outcomes were neither related to region nor population, but rather to the concomitant aspirin dose. "We studied the heck out of that," he said, "because there was a lot of suspicion among our biostatisticians that this was fishing for subgroups." However, they observed the same pattern of effectiveness related to aspirin dose in both Europe and in the United States—people who used the drug with the higher doses of aspirin, which was uncommon in Europe, did not fare as well as those who used the lower doses of aspirin, which was more common for Europeans. "Eventually [ticagrelor] got labeled with, 'Don't use with high doses of aspirin,'" he said. Ultimately, FDA was able to find a solution to this issue because it analyzed the data at the subgroup level, Temple concluded.

DISCUSSION

To begin the broader discussion, moderator Harry Selker, Executive Director of the Institute for Clinical Research and Health Policy Studies at the Tufts Medical Center, solicited comments regarding access to data from trials. If one wishes to re-analyze a previously conducted study to search for latent variables, among other areas of interest, it may be difficult to obtain access to those data, he said, which can be an important factor in how quickly the field advances.

One audience member responded that he would prefer that the researchers who perform major studies release the data collected during the trial for use by other researchers and clinicians after the study is published. In particular, he suggested it would be useful to have "an online calculator so that you can apply the results from the evidence of that trial to the patient before you." How, he asked, can that be made to happen? Joseph Ross, Associate Professor of Biomedical Informatics from Yale University, commented that several clinical trials are now, in fact, being made available by sponsors, manufacturers, the National Institutes of Health (NIH), and others for secondary research purposes. Ross leads the Yale University Open Data Access (YODA) Project, which has partnered with Johnson & Johnson in making clinical trial data available. The YODA Project offers more than 250 clinical trials to which Johnson & Johnson has provided access. There are many additional groups providing access to clinical trial data, Ross said, mentioning www.ClinicalStudyDataRequest.com, pharmaceutical companies such as GlaxoSmithKline and Roche, and NIH's Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), which provides access to data from studies funded by the National Heart, Lung, and Blood Institute. Temple commented that, in addition to data, it is often important to have access to tissue samples, so that additional tests can be performed using those samples when necessary.

Relatedly, Sheldon Greenfield, Executive Co-Director of Health Policy Research Institute at the University of California, Irvine, acknowledged the work described by the panelists, adding that this work should be widely practiced as quickly as possible. "Let's get on with it," he quipped. "In every trial, there should be predictive models of this sophisticated type embedded in the trial, not only to improve the analysis of the trial, but also so we do not have to wait forever and retire before the studies come out." He also suggested enlarging the group of variables examined for their relationship to treatment effect to include patients' personal variables such as comorbidities, functional status, presence of depression, participation in care, and various other social determinants of care. Sherrie Kaplan, Executive Co-Director of Health Policy Research Institute at the University of California, Irvine, compiled a composite variable that captures many of these factors, Greenfield noted. The reason it worked was because her composite captured a latent variable—the patient's ability to respond to treatment. Angus said he strongly endorsed each study having a "multi-attribute risk model of your best understanding of predicting the outcome of disease, even though you might be wrong about mechanism." He would also favor having such a model mandated for every large Phase III clinical trial. "Even if nothing else happened today," he said, "the death of the one-at-a-time subgroup analysis would be great."

Turning to another topic, Kent offered a technical comment about why subgroup analyses based on outcome risk are often easier and more reliable than conventional analyses searching for relative effect modifiers. Obtaining reliable analyses on relative effect modification is difficult for two reasons, he said. First, it is usually the case that little is known about the relative effect modifiers before the trial begins, which leads to "fishing expeditions." Second, trials do not generally have enough statistical power to provide solid data on the relative effect modifiers. This situation causes the forest plots examining relative effects to be unreliable, he stated. "They're unreliable empirically, but also theoretically," Kent continued. "We should anticipate that they're unreliable because they're very underpowered, and the prior information that we have is typically very weak." Throughout the remainder of the session, several participants offered opinions regarding forest plots, with some saying they are useful and others offering caveats about their weaknesses. Rodney Hayward, Professor of the Department of Internal Medicine and the Department of Health Management and Policy at the University of Michigan, suggested that forest plots should often be restricted to the appendices of a paper, thus allowing researchers who are interested to view them and preventing other readers, such as clinicians, from being misled by them.

Ralph Horwitz, Professor Emeritus of Medicine at the Yale School of Medicine, commented that a related problem is that "a lot of the heterogeneity is outside the trial." That is, trials are generally run with relatively narrow inclusion criteria, causing much of the heterogeneity that doctors see in clinical practice to never be included in trials. For example, he continued, the drugs that patients have been taking previously is typically neither reported nor analyzed. "You'd like to see more inclusion of whatever background drugs they were on in the first place?" Temple asked. "I think we [at FDA] are very sympathetic," he continued. "We would like to see the background drugs that people were on kept in at least some of the studies."

Finally, Greenfield mentioned the importance of observational studies. "I'm not talking about big data," he said. "That's a separate topic. I'm talking about intermediate and small data—hundreds, maybe thousands of people." Such studies are becoming more and more common, he said, to the point that they are eclipsing randomized trials. "These observational studies are a rich source of HTE," he explained. "I think we've got to move ever more toward using the data that we have and are able to collect."

3

PATIENT PERSPECTIVES OF THE SIGNIFICANCE OF UNDERSTANDING HETEROGENEOUS TREATMENT EFFECTS

As researchers and clinicians search for the best ways to deal with and take advantage of heterogeneous treatment effects (HTE), they must keep in mind the needs and desires of the consumers of the treatments—that is, the patients. Thus, one workshop session was devoted to exploring the patient perspective on HTE, including what heterogeneity means to patients, how they can benefit from it, and how clinicians and researchers can work with patients to realize the promise of HTE. Much of the information offered in this session was relevant to the patient question: How can I use knowledge about HTE to improve the outcomes that are most important to me?

Points Highlighted by Individual Speakers

- Patients need research that will provide information on how drugs work in people who are not the "average patient." (Concannon, Morgan)
- It would be helpful for patients if studies about a specific condition were carried out in a uniform way so that results could be compared across studies. (Morgan)
- Clinicians need to speak with patients about their treatment priorities. (Stake)
- Patients and practitioners need decision-making tools to help them take advantage of the research into heterogeneous treatment effects. (Davidson)
- Insurance companies will need to develop new ways of making decisions to take advantage of lessons learned from studying heterogeneous treatment effects. (Dubois)

ENGAGING PATIENTS IN DISCUSSIONS ABOUT HETEROGENEOUS TREATMENT EFFECTS

To set the stage for the presentations, Thomas Concannon, Senior Policy Researcher at Tufts University and the RAND Corporation, described his work with the Patient-Centered Outcomes Research Institute (PCORI) to understand patient concerns. PCORI, which funded the workshop, has been working to prepare patients to participate in research. As part of that effort, a team at Tufts University prepared descriptions of the concept of HTE written in a nontechnical manner and then used that material in discussions with three PCORI-funded patient-powered research networks (PPRNs).[1] Each meeting lasted 90 minutes, the last half of which was devoted to structured discussions. During these discussions, participants were given two hypothetical patient scenarios and asked to apply the results of the hypothetical research described in those scenarios, including subgroup analyses, to make decisions regarding their own hypothetical care.

Concannon and his team recorded, transcribed, and then analyzed these discussions. They identified several themes that fell into four major categories: what decisions patients said they made routinely, what kinds of information they seek out in making their decisions, where they go for that information, and how they judge the quality of the information that they discover. Concannon said,

> A single overarching theme was common throughout the discussions about the research, and that is that patients have questions and concerns about the relationship between average results described in clinical research and their own individual case or characteristics. They have concerns about findings from trials that exclude patients like them. And they have concerns about findings couched in averages that obscure how patients like them might fare under treatment.

Many of the patients described dealing with concerns and questions in ways that were "clearly not ideal," Concannon said. One strategy, for instance, was visiting multiple providers until the patient could understand how existing data might be applied in his or her case. Another was treating their own care as an "N-of-1 experiment" until they were able to find a satisfactory treatment; but this can only be done for conditions with symptoms that respond to changes in treatment

[1] Patient-powered research networks (PPRNs) are networks that are organized and operated by patients and their partners. Each PPRN is dedicated to a particular health condition, such as multiple sclerosis or chronic obstructive pulmonary disease. PPRNs work with researchers to shape the research agenda into the disease of interest and to collect health information from their members that can be shared with researchers. Each PPRN is part of PCORI's research network, PCORnet.

in a timely fashion. The patients would also delay making a decision when they felt there was not enough information.

Concannon concluded, "Patient-centered care goes well beyond the usual characteristics of patient preferences and questions of access to care and includes questions around data and the patient-centeredness of data."

Following Concannon's opening remarks, two presenters and two responders spoke about the patient perspective. A key theme of their presentations was that the traditional relationship patients have had with the health care system will not be sufficient to meet the demands and the opportunities of HTE. The presenters offered a variety of suggestions for the necessary changes.

THE PROBLEM WITH TREATMENTS AIMED AT THE "AVERAGE PATIENT"

"Traditional research failed to help me when I was first diagnosed with MS [multiple sclerosis]," said Seth Morgan, a neurologist and a fellow of the American Academy of Neurology. His experience as a patient in the medical system convinced him of the importance of the current movement to understand HTE and provide patient-centered care.

Fourteen years ago, Morgan said, he was a neurologist in private practice when "I had the unique experience of diagnosing myself with multiple sclerosis." Following his diagnosis, he continued practicing as a neurologist for a couple years until he felt he was no longer performing adequately. He has since been a patient advocate for the National Multiple Sclerosis Society and for people with disabilities generally.

When he and his doctor, a fellow neurologist, first set out to find a treatment plan, Morgan said, he quickly ran into a conundrum. "The premise of traditional research," he explained,

> is to put a treatment at the center of consideration and decide, Is this treatment helpful for an average patient? Trouble is, there aren't very many average patients out there, and I, like most people, am not an average patient. So traditional research could not answer the question, the basic question that everyone wants to know, including patients and their caregivers, and that is, What is the treatment that is most likely to help me or my patient with their specific issues?

When Morgan was first diagnosed with MS, the only available treatments were injectable medications. After researching the options, he and his doctor "by gestalt" determined what they felt was probably the strongest medication on the market. "That was the best we could do—a best guess." One problem Morgan had with the

medication was his phobia of needles. "I did not like needles." As his only choice, he stayed on the medication, giving himself an injection every other day for 2.5 years. But he never got used to the needles, and after those 2.5 years, he simply could not do it any longer. "I went to my physician and said, 'I know the data [are] against us, but I'm stopping this medication. I don't care if it's going to cause progression of the disease or not. I just can't be on injectable medications anymore.'" By chance, GILENYA®, the first oral medication for MS, was being studied in clinical trials at the time, and he was able to start taking it instead. After about 1 month on the new medication, Morgan said his wife turned to him and said, "You're back." What do you mean? he asked her. And she explained, "You were cognitively affected. You were duller than you normally were. And now you've cleared up."

Neither Morgan nor his doctor noticed that the injectable medication caused subtle cognitive issues—a side effect that Morgan suspects might have been anticipated if less attention had been paid to the "average patient" and more paid to him as an individual. He was an older individual who developed MS and also had strong family histories of Alzheimer's disease in both parents. If the proper studies had been conducted, researchers might have recognized that this particular drug could cause cognitive issues in members of a vulnerable subpopulation—and he might have avoided 2.5 years of dulled cognitive function.

Today, he noted, there are multiple treatments for MS. Yet, there is still no way to provide doctors and patients any direction on which medication should be tried first or which medication is most promising for a particular individual patient. A big part of the problem, he said, is that it is impossible to compare the outcomes of different studies "because of the different parameters and specifications that were delineated by the individual pieces of research." Ideally, studies would be conducted with uniform selection criteria and comparable treatment regimes and analyses to enable valid comparisons among treatments.

Even better, Morgan continued, would be trials that compared the performance of different drugs in various subgroups of patients. For each subgroup, the trial would ask, Which medication has the least likelihood of causing side effects and the greatest likelihood of providing benefits? "Of course," he said, "the trouble is that you're not going to get drug companies to fund those because they don't want to have head-to-head studies against a different drug in case they happen to come out on the short end, and they have no control over the way in which patients are selected."

Still, he believes the situation is improving.

> What I feel currently is happening—and I think is a good thing—is that research is going through a transition, and they're looking at the question of what is going

> to help this particular individual based on [his or her] clinical issues. I think that's an important paradigm change—how to determine what individual patients, or real patients, should be treated with when they present with a problem. No one is currently able to answer that question. No one could answer it 14 years ago when I first started medication.

"It's a real big problem," he concluded, "and I think it's important that we continue this movement toward patient-oriented treatment because we need to understand the situation of each individual as best we can and figure out what subgroup of individuals is most likely to benefit or not benefit from a given treatment."

TAKING PATIENT PREFERENCES INTO ACCOUNT

In the next presentation, Christine Stake, Research Operation Manager of the Ann & Robert H. Lurie Children's Hospital of Chicago, described two of her experiences with the health care system that indicated the importance of doctors communicating with their patients and understanding what is important to them in a treatment.

The first experience she described was with severe osteoarthritis in her hip. As a child, she had hip dysplasia. And by the time she was in her early 30s, she was in severe pain—to the point that she had to crawl up her stairs. When Stake went to a doctor, she was told she needed a hip replacement, but she was "too young." Her only other options were physical therapy and pain medication, both of which stopped working within 6 months. So, again, she went to the doctor, only to hear same thing: "You need a hip replacement, you're too young."

When she asked for further explanation and for the specific contraindications, Stake was told that, because of her young age, she would likely need yet another hip replacement in her lifetime. When she asked for evidence, however, her doctor could not provide any—there were no studies of hip replacements in young people because such operations were rare. Consequently, there was no way for her—or her doctors—to know how long a hip replacement was likely to last in a young person like herself.

Throughout her search, she saw several doctors and heard the same answer from each: No hip replacement because you are too young. Stake was frustrated for a couple reasons. First, the doctors were making a decision without any research to guide that decision. Related research had shown that, in the elderly population, hip replacements are successful, she said. "For the average patient, the outcomes were incredibly high; but this surgery was not being allowed

to be offered to me." Her other problem was she felt the doctors were acting paternalistically. She understood, however, they thought the likelihood of her having to replace her hip twice was too high. In short, they were basing the decision on what they thought was the most important factor: risk of a repeat surgery. Yet, they never communicated with her about what *she* thought was the most important factor or about her concerns about quality of life. Was she willing to make the trade-off of someone in her 30s trying to have a life versus the risk of possibly needing another hip replacement later in life? No one had that conversation with her. Finally, Stake said, she found a doctor who talked with her about a hip replacement in terms of these trade-offs. Jointly, she and the doctor decided that she would indeed get a hip replacement.

"As a patient," Stake explained, "I don't walk in and expect a doctor to say, 'Based on you as an individual, this is the right decision for you.' But I do expect a conversation—to say, 'These are your risks, these are your benefits, let's make the best educated decision we can.' … We really have to talk with patients about what's important to them." For her, Stake said, the hip replacement was the right choice. "I had that surgery. The minute I woke up from surgery, my joint pain was gone." Afterward, she returned to school and completed her doctorate, for which her research focused on patient decision making.

That experience reinforced the lessons she learned nearly a decade earlier when, at 24 years old, doctors found a lump in her throat. They did not expect it to be cancer, she said. Indeed, the doctors told her it was highly unlikely to be cancer—but when the tests came back, it was in fact cancer. She needed her thyroid removed. The question for her was, What sort of follow-up radiation treatment would she choose? She could choose to have either a high dose or low dose of radioactive iodine to clear up any remaining cancer cells in her throat. Normally, she was told, a patient would be given the high dose, but there was a chance that the high dose might affect her fertility. Which did she want?

Once again, the problem was, as she noted, the lack of evidence to support whichever decision she made. All she knew was the low dose meant a higher risk of her cancer returning, while a high dose was more of a threat to her fertility. She chose the lower dose. "I made the decision based more on quality of life because I didn't have the evidence to make that decision," she explained.

One lesson that Stake drew from her experiences is the importance of getting information that helps outlier patients make decisions about treatment. She said,

> We understand as patients that you have to look at averages. We all know the bell curve, we all know the two standard deviations. But how can we better do

> studies to capture those two-standard-deviations patients? When you go talk to your patient and say, "You qualify for this medication" or "This is a good medication for you," you can feel good about that. But what about that patient where you say, "You don't qualify for that clinical trial" or "You don't qualify for this"? Those are the ones who really need to have more of that shared decision-making model and more of that patient-centered care conversation because evidence doesn't lead them to an easier decision.

Stake concluded by challenging those in the audience to observe things differently. "How do we assess patients on that continuum and not always ignore the two standard deviations or those outliers? And how can we maximize those opportunities to have those conversations with patients?"

PROVIDING PATIENTS WITH DECISION-MAKING TOOLS

In her response to the presentations, Karina Davidson, Professor of Behavioral Medicine at Columbia University, focused on the need to provide patients with the proper decision-making tools to help them figure out how to deal with information about treatment heterogeneity.

"I agree that heterogeneity is a big barrier—or opportunity—for the practice of our clinical work in the next generation," she began. Indeed, she said, as a clinical psychologist she is used to dealing with patients who tend to need personalized solutions. The depressed, the anxious, the schizophrenic, the addicted, the obese, the smokers who are trying to quit—for none of these patients is there any sort of "magic pill" that has been clinically shown to help everyone with that problem.

"My whole life," she said, "has been thinking about the heterogeneity of treatment effect and the fact that we ask patients to deal with this completely weird counterfactual: What will your life be like on this lifelong drug that is going to cause you palsy, cognitive decline, and all these side effects but clear up reality-versus-not? And how can they possibly imagine that?" Furthermore, the evidence available to clinical psychologists may or may not be relevant to a particular patient, and its applicability will only become known after the treatment is done.

With that, she moved to what she described as her main point—that evidence-based medicine as originally described by David Sackett is actually evidence-based practice, not medicine. "He abhorred the idea of recipe medicine," she said, "and implored us to think through, empirically and with evidence, how we incorporate patient preferences, values, and their preferred outcome into the art of the practice.

But that piece was hard, so we worked on—I think, importantly—the randomized controlled trial data." But the result of that, she said, is that neither patients nor practitioners ended up getting the evidence they needed in the way they needed it in order to apply it quickly and efficiently.

"We could do better than that in heterogeneity of treatment effect," she said. If the research is done, but the needs of the practitioners and patients are not taken into account, then the job is only half done, she said. It is important, she said, that medical researchers and others in the health care community have a conversation about the sorts of dissemination and implementation tools that will help practitioners and patients understand and apply what is being learned about HTE. "If we can come up with ways that are innovative and new and join up with the people in dissemination and implementation," she said, "we might have some really great ways that we can start … figuring out the best practices for informing patients and practitioners about when heterogeneity of treatment effect is expected and what the best course is for a patient."

INSURERS AND HETEROGENEITY

It is not just clinicians and patients who will need to learn to deal effectively with treatment heterogeneity, said Robert Dubois, Chief Science Officer and Executive Vice President of the National Pharmaceutical Council. Health insurance companies and other payers must adapt, as well. In his response to the panel presentations, he spoke about the challenge that HTE poses to health insurance companies.

The development of new medical treatments depends on a virtuous cycle, he said: There is a need, a drug is developed that meets that need, the drug gets used, it gets paid for, and then some of those dollars are used to pay for developing new treatments. For the virtuous cycle to succeed, Dubois said, the relevant patients have to perceive that a treatment is beneficial to them, doctors have to recognize it as the preferred therapy for those patients, and, crucially, insurance companies have to pay for that therapy.

How do insurance companies decide which treatments to cover? Increasingly, Dubois said, their decisions depend on various sorts of value frameworks, and for the most part those value frameworks base their determination of value on the average patient. There are some exceptions, he noted. When a genetic test can predict whether a patient will respond well to a particular treatment, for example, insurance companies may value the treatment differently for different drugs.

The vast majority of decisions, however, do not take into account many of the factors that are important to patients, he said. Consider the choice of a cancer

therapy. Some regimens are very complex for the patient to deal with, while others are simple and straightforward. Some require injections, others do not. Some have an effect on one's quality of life, others affect quantity of life. And different patients value these things differently. One patient may want a straightforward regimen that will not ruin his quality of life, and he will accept that he may not live quite as long as he could; while another may be willing to put up with any sort of regimen, any sort of side effects, because she really wants to live as long as possible. Yet, when an insurance company does a value determination for this sort of cancer drug, it gets a single answer: Its value is X dollars.

"The problem," Dubois said, "is that patients do not fit into simple dichotomous categories; and if you have 3, 4, 5, 10 types of outcomes that are important, then you have to take those into account. And, unfortunately, we're increasingly seeing value frameworks that cannot handle that."

This is not just a theoretical problem, he said. New York State just carried out a review of a new cystic fibrosis drug based on cost per outcome, and it decided that the drug did not meet its threshold and would not be paid for. It was a one-size-fits-all assessment of that cystic fibrosis drug. However, Dubois said, it is very likely that because different patients would have different preferences and experience different outcomes with that drug, some patients would find that particular drug to be very valuable for them, perhaps more valuable than any other drugs that are on the market. "Those patients will not have access to the drug," he said. "And we go from a virtuous cycle to a vicious cycle because if drugs come to market and they're not being valued correctly for different subgroups of patients, patients won't get it, it won't be paid for, and then new drugs won't get developed."

The take-away message, Dubois concluded, is that in the era of recognizing HTE, it is not just clinicians and patients who will need to come up with new ways to make decisions. The payers, too, will need to develop new ways of determining value and making decisions about which treatments will be paid for and which will not.

DISCUSSION

In the discussion session following the presentations, Richard Willke, Chief Science Officer of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), spoke about how U.S. insurance companies might deal with treatment effect heterogeneity. Generally speaking, he said, most U.S. insurance companies do not have a one-size-fits-all approach to paying for medical treatment. Most have "utilization management," which attempts to draw

distinctions among patients and make reimbursement decisions based on some individual factors. The insurance companies are an important audience for the topics being discussed at the workshop, he said. "They'd like to know better. We maybe don't give them enough credit for trying to do the right thing." Their actuaries are pretty sophisticated, he said, but they need the right sorts of evidence to make those decisions. "We ought to think about that audience as well."

Dubois agreed, and he said that it may not be as difficult as one might first think for insurance companies to make decisions that take both HTE and patient preferences into account. Although, in theory, patients could have a wide variety of sets of preferences that insurance companies would have to deal with, Dubois said he believes there may be just a few groupings that would cover most patients. One group of patients, for example, puts more focus on quality of life, while another wants treatments that will extend life as much as possible, and there are perhaps two or three other subgroups. Once the subgroups and their preferences are determined, it will be necessary to come up with utility scores for different options, with each subgroup having different utility scores. Then by combining the utility scores with data on the likelihood of various outcomes, it would be possible to calculate the expected benefit of various treatment options for each of the subgroups. It has not been done before, Dubois noted, but it could be a workable approach.

Naomi Aronson, Executive Director of Clinical Evaluation, Innovation, and Policy of Blue Cross Blue Shield, also addressed the question of the role that value plays in determining what insurance companies will pay for treatments. "There is a lot of discussion of value now," she said. However, she added, insurance companies are contractually obligated to determine payments according to a medical necessity model, not a value model. Many clinicians are concerned with how that model is skewing payment decisions, she said, because "low-value cancer care is driving out very high-value treatments for, say, inflammatory arthritis." In response, there is a push toward adopting value-based payments, with insurance companies paying less for procedures and individual services and more for episodes of care. "There is a commitment to drive payment in that direction," she said. "And it is happening." It probably will not be easy, however. In order to move toward value-based payment, insurance companies must determine who will benefit from procedures and who will not, "which is exactly the heterogeneity we're talking about here."

Turning to a different topic, Frank Davidoff, Editor-in-Chief (Emeritus) of the *Annals of Internal Medicine*, offered a comment about medical care being a combination of both a production model and a service model. He spoke of a

movement he described as "being developed largely under the leadership of Paul Batalden" of the Institute for Healthcare Improvement.

"Basically, the model they have begun looking at is the model in business of what they call the production model and comparing that with the service model," Davidoff said. "In a production model, you produce a product and you send it out into the marketplace, and it gets used," he explained. It is, in a sense, a one-way transaction. By contrast, the service model is "a reciprocal, repetitive, cyclic, ongoing kind of process which is continuously adapting and changing." Examples of the service model include everything from house painting to lawyers to social work.

"Medicine has the disadvantage, it seems to me, of sharing the qualities of both," Davidoff said. But in some respects, he suggested, we are trying to fit a service industry in a production model. This leads to much of this discomfort and tension in medicine that had been talked about at the workshop, he suggested.

4

NEW METHODS FOR THE PREDICTION OF TREATMENT BENEFIT AND MODEL EVALUATION

If researchers, patients, clinicians, and payers are to deal effectively with heterogeneous treatment effects (HTE), it will be necessary to develop new methods and models for predicting the benefits and the risks for individuals that are posed by various treatment options. This session looked at various approaches that are available to—or that can be developed by—medical researchers for studying and predicting HTE, focusing on risk modeling based on both genetic data and statistical tools. Much of the information offered in this session was relevant to the patient question: Given my personal characteristics, conditions, and preferences, what should I expect will happen to me?

Points Highlighted by Individual Speakers

- Despite years of effort, polygenic risk scores have not yet proved particularly effective in providing clinically useful predictions of disease likelihood. (Janssens)
- A great deal of theoretical work has been done on developing machine learning tools that could be used to analyze heterogeneous treatment effects, but additional efforts are necessary to get them ready for clinical application. (Li)
- Clinical decision making inevitably takes place at the level of the individual, but the validation of clinical decision-making tools takes place at the level of the strata, and performance is judged at the population level. (Heagerty)

POLYGENIC RISK SCORES

In the first presentation, A. Cecile J.W. Janssens, Professor of Epidemiology at the Emory University Rollins School of Public Health offered a brief history of efforts to use multiple genes to predict the risk of individuals developing common diseases such as breast cancer and heart disease. The idea is not new, she began. It can be traced back at least 20 years. Little progress, however, has been made in the area since that time.

In the late 1990s, medical researchers began to talk seriously about using genetic information to predict susceptibility to common diseases. In 2002, a paper (Pharoah et al., 2002) published in *Nature Genetics* titled "Polygenic Susceptibility to Breast Cancer and Implications for Prevention" was one of the first to discuss polygenic susceptibility, she said, and it had one of the first mentions of risk distributions. Specifically, that paper described (1) how certain women had risk distributions for breast cancer that were different from the risk distribution for the general population, and (2) how polygenic risk could be applied in health care to make mammography screening more cost-effective, by identifying those women at greater risk who should be given more frequent screenings. This concept, she explained, is the basic idea underlying polygenic research—that is, using polygenic risk scores to identify high-risk groups for targeted intervention and, more generally, to apply different interventions for different risk groups.

The following year, in 2003, a paper (Yang et al., 2003) appeared in *The American Journal of Human Genetics* that was the first to describe how multiple genes can be combined to predict risk by using regression analysis. The study focused on posterior risk for carriers of one or more multiple risk alleles, and the particular alleles it examined had strong per-allele effects by today's standards, with risk ratios between 1.5 and 3.5. The researchers concluded, Janssens said, that it is a promising endeavor to combine multiple variants into a risk score.

In response to that 2003 paper, Janssens and several colleagues, including Ewout Steyerberg, the panel's moderator, wrote a letter to the editor of the journal arguing that if one wanted to evaluate a polygenic risk score—though it was not called that at the time—it was necessary to include not only high-risk people, but all people, including those who were not carriers of the risk alleles (Janssens et al., 2004). Because the paper's authors had not done that, the letter contended, the usefulness of the score was difficult to evaluate.

In that same letter, Janssens and her colleagues recommended using a well-established measure called the area under the receiver operating curve, or AUC, to evaluate how predictive a risk score is. In essence, she explained, AUC is an indication of the separation between the risk distributions of those who develop

a disease and those who do not. A small AUC indicates a large overlap among the risk distributions, meaning that the risk score does little to differentiate those who develop a disease from those who do not, while a large AUC indicates that the risk distributions are noticeably distinct. In particular, Janssens said, an AUC indicates how well a measure will be at identifying people who will develop a disease at the cost of how many false positives that might result. "When you want to select the highest-risk group and your AUC is higher, there are more people who will develop the disease in your selected high-risk group, whereas when your AUC is very low, there are many people who will not develop the disease in your high-risk selection. It's just not as good as you think it is."

Then Janssens used AUC to discuss the discriminative ability of some polygenic risk scores that were developed a decade ago, before the use of genome-wide association studies (GWASs) became common. For example, a series of studies of type 2 diabetes scores ended up with AUCs that were between 0.55 and 0.60—too small to have much of a separation among the risk distributions. On the other hand, studies of age-related macular degeneration (AMD) and hypertriglyceridemia scores produced AUCs of 0.80—large enough for a substantial significant separation among the distributions.

Generally speaking, she said, the polygenic risk scores with large AUCs rely on common variants with strong effects on a person's risk. She compared, for example, the gene variants that went into the two risk scores, one for type 2 diabetes and one for hypertriglyceridemia. The type 2 diabetes risk score, which produced an AUC of 0.60, relied on 18 gene variants, none of them with a risk ratio larger than 1.36 and half of them 1.10 or less. By contrast, the hypertriglyceridemia risk score used only seven variants, but they all had risk ratios between 2.10 and 7.36 (Lango et al., 2008; Wang et al., 2008).

One possible way to use genetic factors to predict risk would be to combine them with traditional risk factors and see if the additional genetic information increases the predictiveness. In 2008, she and a colleague reviewed a dozen papers that took that approach (Janssens and van Duijin, 2008), and they found that the addition of information about genetic variants to the traditional risk factors increased the AUC by 0.06 or less and, in most of those studies, by 0.02 or less. In other words, the added genetic information did very little to improve the ability to distinguish those who would develop a disease from those who would not. The genetic information, while interesting, did not have clear clinical implications in terms of modifying how a doctor would approach a particular patient.

Summing up the experiences from that time period, Janssens commented that researchers were still learning how to use the relevant methods, and relatively few genes had been identified, so perhaps no great impact should have been expected.

Other shortcomings included that researchers paid little attention to what the clinical uses of their work might be, they used non-representative populations, they provided no relevant comparisons with clinical risk models, and they tended to rely on p-values instead of AUC—they were just not paying attention to actual improvements in predictions. To be fair, she said, the researchers themselves reported that the methods had limited predictive power, but people believed that once the whole human genome became accessible, the predictive power should improve. "The door was wide open for GWAS to deliver," she said. "The future was still bright."

Indeed, in the past year, she has seen an increased interest in polygenic risk scores. She offered a sampling of headlines predicting that genetic scores would make it possible to predict such things as heart disease, breast cancer, Alzheimer's disease, and even intelligence and reading ability. "You ask yourself, What has changed in the meantime?" she said. "Have all these genes that have been discovered delivered so much?" The answer is no, she said. "The risk distributions are still largely overlapping for the same diseases." She then showed a recent paper using polygenic risk score to predict ovarian cancer. The risk score relied on 96 single-nucleotide polymorphisms (SNPs), and yet the AUC was only 0.60. And a genome-wide risk score that used 6.6 million variants to predict coronary disease had an AUC of only 0.64 (Khera et al., 2018). Using the entire genome to generate a risk score does not do much more—at least in this case—than the studies a decade ago that relied on just a few genes or, indeed, risk scores generated from traditional risk factors.

Why then, she asked, are the people performing these studies claiming to be predicting something? The answer, she said, is that what geneticists mean by prediction is very different from what clinicians expect from prediction. She quoted a recent article by Carl Zimmer that appeared in *The Atlantic*:

> When geneticists use the word *prediction*, they give it a different meaning than the rest of us do. We usually think of predictions as accurate forecasts for particular situations.… Geneticists are a lot more forgiving about predictions. (Zimmer, 2018)

When geneticists speak of "prediction," Janssens further explained, they are generally referring to weak effects with little or no practical or clinical significance because the scores do little to differentiate among groups of people.

Concluding, Janssens noted that the shortcomings of polygenic risk studies today are the same as they were a decade ago. Generally, no consideration is given to the intended use, so there is no way to tell whether the predictive performance

is sufficient. The risk thresholds, if they are chosen at all, are based on little to nothing. There is no calibration, no validation, and no appropriate comparison with clinical models. "I think you can hear the frustration in my voice," she said. "A lot of people are calculating these polygenic risk scores, but they have no clue what can be done with them." Of course, there are exceptions, but this is her overall impression of the field. The problem is that genetic researchers are generating a lot of research results without any framework to understand whether the results will ever be relevant.

In closing, she quoted an article published in *The New England Journal of Medicine*:

> In our rush to fit medicine with the genetic mantle, we are losing sight of other possibilities for improving the public health. Differences in social structure, lifestyle, and environment account for much larger proportions of disease than genetic differences. Although we do not contend that the genetic mantle is as imperceptible as the emperor's new clothes were, it is not made of the silks and ermines that some claim it to be. Those who make medical and science policies in the next decade would do well to see beyond the hype. (Holtzman and Marteau, 2000)

"We are long past that 'next decade,'" she stated, "but I think that it still applies today."

Frank Harrell, Professor of Biostatistics at Vanderbilt University offered several comments on Janssens's presentation. He began by referring to it as appropriately pessimistic, commenting that genetics had not had—with some exceptions—a good track record in being predictive of outcomes. Referring to a paper published by other researchers at Emory (McGrath et al., 2013) and whose results were featured in *The New York Times* (Friedman, 2015), Harrell said, "We get involved in sexy things, and we forget to do the simple things." The findings in the paper that were picked up by the press related to using brain imaging to determine whether a patient would respond better to psychotherapy or drug therapy with antidepressants. Harrell said it was unlikely that part of the study would be replicated, but hiding in one of the paper's tables was a result that, while unsexy, had useful clinical implications: Patients with high anxiety levels responded very differently to psychotherapy than to drug therapy. Precision medicine faces a similar issue, he postulated, predicting that if people in the field are not careful they will end up excelling in "precision capitalism," that is, receiving plenty of grants, but not so much in improving public health. He referred to Janssens's presentation as a "wakeup call" to that possibility.

PROMISE OF MACHINE LEARNING

With the availability of high-speed, high-powered computers, it has become possible not only to quickly analyze large amounts of data, but also to analyze such data in ways that were not feasible before. One such method is machine learning, through which computers search for patterns in data rather than analyzing the data in a predetermined way. Machine learning makes it possible to spot correlations in data that may have never been considered by the machine's human operators and that may not even seem to make sense at first.

In the session's second presentation, Fan Li, Associate Professor of Statistical Science at Duke University described machine learning and discussed how it might be put to work analyzing HTE. Right now, "HTE and machine learning are two buzzwords in comparative effectiveness research," Li said. The use of machine learning to explore HTE has become a hot topic, especially in cases when there are large amounts of data. "The central goal from a statistical standpoint," she said, "is the same as the traditional regression methods: to accurately learn from the data what the outcome function is, given the covariates and the treatment variable. The goal is the same; it's just some new analytical methods." These machine learning methods are generally more flexible and adaptive than the traditional methods, she added, but they are not a panacea. They have their own limitations.

For the duration of her presentation, Li focused on four specific machine learning approaches that are popular methods for application to HTE:

- Penalized regression (e.g., least absolute shrinkage and selection operator [LASSO], elastic net regularization);
- Regression tree-based methods (e.g., classification and regression trees [CART], random forests);
- Bayesian nonparametric models (e.g., Gaussian processes, Bayesian trees); and
- Ensemble learners (e.g., boosting, forests).

"These are all supervised learning methods," and in their original form they are used for prediction, not estimation, she said. In statistics, there is a subtle difference between prediction and estimation—in that, prediction uses observed data on one set of variables to guess at the value of a different variable, and estimation uses observed data to guess at the true value of an underlying parameter. Applying these supervised learning methods to HTE requires that they be modified somewhat. The basic problem, she explained, is that you want to see how a given individual will do under a certain treatment condition versus how that same person would

do under a different treatment condition. "You never see that at the individual level," she said. "That is the fundamental problem of HTE." Over the past 25 years, however, there has been a great deal of work to adapt machine learning models to this sort of analysis, and there are now a variety of machine learning techniques that can be used to predict HTE.

The first of her machine learning approaches—penalized regression—has been used a great deal in work with HTE, Li said. "You do not penalize regression," she explained. "You penalize the complexity of regression." A HTE regression analysis is essentially examining the interactions between the treatment variable and multiple covariates. But many of those interactions are not particularly important in understanding the treatment effects, and you can throw those out without losing much analytical power. "This type of regression essentially has a way to select the most important and meaningful interactions," Li said—and thus avoid overfitting.

Next, she explained regression trees. They get their name because the covariate space is partitioned into subgroups that are referred to as "leaves." The analysis proceeds by predicting responses in each leaf using the sample mean in that region. At each step, researchers review the variables and decide whether and where to split the sample into different leaves, with the tree's complexity growing as more leaves are put in place. Cross-validation is used to make decisions about the tree's complexity—in essence, how deep the tree's branching will be—and about how parts of the tree can be "pruned" to reduce its complexity. The technique was originally developed for prediction, Li said, but in 2016, it was modified for use with HTE in what Li referred to as a "landmark paper" (Athey and Imbens, 2016). It is just one of what Li said are "probably hundreds of papers written in the past few years about doing this sort of analysis."

There are several advantages to the regression tree approach. There is a large selection of available software that can be used to implement it, for instance. It is fast, and it is easy to interpret. Additionally, while a single tree may be a "weak learner," one can average over a collection of trees—called a forest—to improve the estimates. The disadvantages of the regression tree approach, she said, are that the trees tend to underestimate uncertainties, and that the structure of the trees lacks the flexibility of some of the other methods.

The third approach is Bayesian nonparametric models. Li described two types of Bayesian approaches: Bayesian trees (which are similar to a regression tree except that they are implemented under the Bayesian paradigm, using prior knowledge) and Gaussian process models. Generally speaking, Bayes' theorem is an approach to calculating probabilities that begins with a certain amount of prior knowledge—unlike traditional statistics, for which the calculations depend

completely on the data that are collected—and then modifies the probabilities as more and more data come in.

Discussing the pros and cons of Bayesian nonparametric methods, Li said that the pros include the fact that you can incorporate prior knowledge into the model, the model quantifies uncertainties automatically, it works well with small samples, and it is "elegant," which makes it appealing to mathematicians. One of its biggest cons is that these approaches are difficult to scale—as the amount of data grows, the required computational resources increase rapidly. Furthermore, these approaches can be difficult to explain to a lay audience, software for implementation is limited, and having to choose and justify prior distributions before getting started is an additional complication.

The fourth type of approach Li described is ensemble learners. In any of the first three approaches, she said, a single model can be a weak learner. "The performance might not be good. So, the idea is to make a bunch of models and combine them. That's where the name 'ensemble learning' comes from." A forest is one type of ensemble—in this case, when you assemble a bunch of trees—but the same idea can be applied to other models, as well.

Summarizing her presentation, Li highlighted several points. First, there has been a great deal of work recently on machine learning theory, both in statistics and in economics, but there is still much to learn about applying it to health statistics and HTE. There are still relatively few applications that have been developed, and "there's a huge gap between theory and practice here." Thus, more translational work is needed. In the past couple years, researchers have developed several new methods for use with HTE that are being offered to those in the health field. However, Li said, there is little that is known about the "empirical, comparative performance of those models," and that issue needs to be addressed. Software development will also be key to using machine learning to analyze HTE. Many of the methods she described have available software that no one is using because people in the HTE field are not yet familiar with it. A change will require effective collaboration between the methodologists—the statisticians, the machine learning researchers—and the clinical researchers, Li said. Finally, she said, "One must organically fuse traditional statistical tools and machine learning to reach better comparative effectiveness research." It will not work simply to "put one upon the other" because there are a number of subtleties about combining the traditional statistical tools with machine learning that must be worked out.

Frank Harrell also commented on Li's presentation. One of her key messages, he said, was that machine learning methods are very flexible, but they are also "data hungry." Referring to a paper by Ewout Steyerberg and colleagues, he noted that machine learning analyses could require as much as 10 times the sample size that

traditional regression analyses require (Steyerberg et al., 2014). Regarding the various models that Li described, Harrell said, "To me, the method that has the most promise of anything in this space is ordinary Bayesian parametric models" because of the possibility of using expert opinion to guide both the studies and the data collection.

METHODOLOGICAL ISSUES RELATED TO PREDICTIVE SCORES

Patrick Heagerty, Chair of Biostatistics at the University of Washington School of Public Health, discussed some technical issues related to developing and evaluating predictive scores. "What can we say from the data that we collect and the analyses that we produce?" he asked. "I want to emphasize limitations in our ability to make specific statements."

Generally speaking, Heagerty said, there are three levels at which one can analyze risk. The first is the individual patient level, for which one ideally would like to make predictions. Generally, however, such predictions are not feasible because it is impossible to observe outcomes under counterfactual treatment conditions in an individual. Consequently, what happens is that one makes predictions at the level of a stratum—a group of individuals with similar characteristics. This is the second level. The third is the population level, which is the level that is most important when talking about maximizing performance. He then spoke briefly about the recent work in prediction in the field of statistics, emphasizing at which level (i.e., individual, stratum, population) the work was being carried out. He also noted the work of Susan Murphy, who focused on identifying rules for treatment programs at the individual level followed by measuring the effects of those rules at the population level (Qian and Murphy, 2011).

One quantity of interest is referred to by statisticians simply as "the value," he said; yet, it is actually the mean population outcome under the targeted treatment. Two approaches are used to determine the population mean. One is Q-learning, in which one estimates the mean under both treatment and control and compares them; one can then recommend treatment based on whichever mean is better—a rule that will ultimately lead to a better population mean, he said. The other is outcome-weighted learning, in which one skips the outcome model and directly makes a prediction rule that optimizes performance at the population level. In both approaches, the focus is on decisions at the individual level but performance at the population level. "That's the first message that I really want to emphasize." He continued, "Many contemporary methods consider action at the individual level but still measure performance at the population level. I think it is appropriate, but there is a disconnect."

Haggerty next arrived at his second main point: Most attempts to define "patients who benefit" are unreliable because they rest on outcome-based definitions that are fundamentally not measurable, owing to the fundamental problem of causal inference for individuals. Because researchers cannot observe the outcome under an alternative treatment for a given patient, "it's very difficult to migrate some of the tools we've used traditionally for diagnostic and prognostic methods. We heard this said at least twice today, and I'll just make it a little more formal." The ultimate goal, he noted, is to find predictive markers, or markers that can be used to guide treatment for individuals. But what, he asked, does it even mean to talk about individual benefit? "What do we mean when we say 'patients who benefit from treatment'? Can you label those patients for me, the patient [who] will benefit from treatment? And the fundamental answer is no." Why? Determining which individuals will benefit from treatment requires that we know how that patient would fare if treated and if not treated—which is impossible since a patient cannot be both treated and not treated.

To formalize this notion, Heagerty presented a figure listing the various possible outcomes for a patient undergoing treatment (see Figure 4-1). In this simplified world, outcomes are either positive or negative, with no gradations. There are four possibilities for how an individual patient will respond: positively if left untreated and positively if treated (i.e., neutral between treatment and no treatment); positively if left untreated and negatively if treated (i.e., worse outcome with treatment); negatively if untreated and positively if treated (i.e., benefit from treatment); and negatively if untreated and negatively if treated (i.e., neutral). Ideally, a clinician would like to know in which row of the figure a given patient fits to inform treatment decisions. The fundamental problem, however, is straightforward, Heagerty said. "We can see data for untreated people, whether they get better or not or whether they do well or not. We can see it for treated people, as well." But, as he previously mentioned, you do not know how each untreated individual would have done with treatment or how each would have done without treatment.

These four categories can also be viewed as "principal strata." In this regard, Heagerty mentioned the principle stratification framework, introduced by Constantine Frangakis and Donald Rubin (2002), as one approach to estimating the causal effects of treatment. Heagerty alluded to the controversies over the appropriate uses of this causal inference framework, citing contradictory papers (i.e., Janes et al., 2015; Simon, 2015), which were published in the same issue of *Journal of the National Cancer Institute*.

Notably, it is possible to sidestep the limitation of unobservable counterfactual outcomes, he said—particularly when the outcome of the treated or the untreated

Patient type	Y(0)	Y(1)
(neutral +)	+	+
Worse	+	-
Benefit	-	+
(neutral -)	-	-

- Y(0) = untreated outcome
- Y(1) = treated outcome

FIGURE 4-1 | Potential outcomes for a patient undergoing a medical treatment.
SOURCE: Patrick Heagerty presentation on May 31, 2018.

condition is relatively uniform. Suppose, for example, that you limit yourselves to a group of patients who will clearly have a negative outcome without treatment. "This may be some oncology setting where we think that without treatment people will do poorly," he said. In this case it is only necessary to know how a patient will do with treatment. The critical question then becomes, Is there a biomarker or some other measure that separates people who will do well under treatment from those who will do poorly? "In this one setting we can start to migrate tools for classification or prediction," he said. A second situation in which the problem can be sidestepped is when the treatment is expected to always—or almost always—have a positive outcome. In this scenario, the prediction issue is simplified to an issue of prognosis: Who is at greatest risk if left untreated, and who will likely be fine without treatment? In that case, there are several tools that can be brought to bear (Steyerberg et al., 2010).

Furthermore, there are situations in which one can measure both conditions—to see how a patient will fare with and without treatment. Describing the "N-of-1" trial approach, Heagerty suggested that patients with pain, for instance, can be treated first with a pain medication for a period of time and then with a placebo, or vice versa, and the outcomes compared. This is not a perfect solution, however, as other variables may change over the time periods of treatment. "There are still identifiability problems in this space," Heagerty explained. "It also invites other questions, like, Is that really the goal, whether my one outcome is better than my other one outcome? Or is it really my mean outcome, repeatedly treating me one way and repeatedly treating me another way?" That said, Heagerty reiterated his second main point, "I think attempts to define 'patients who benefit' … can lead to an outcome-based definition that just is not measurable. This is a fundamental problem about talking about the performance of classifiers for [whom] should get treated."

Regarding his third point, Heagerty described the use of scores. In particular, he spoke of the score as the difference between treated risk and untreated risk.

"For given characteristics, what's the difference if I'm treated as compared with if I'm not treated?" Noting that many people in the workshop discussed the importance of prognosis—knowing what is going to happen in an untreated patient—he emphasized the importance of prediction scores, that is, predicting what will happen to a treated patient. "I think the fundamental goal is to try to learn that predictive score, the benefit that would be assigned to a given patient," he said. "I want to push us to say, 'Yes, it's important to look at baseline, untreated, prognostic risk, but let's at least start to try to get scores that measure the ensemble, aggregate expected benefit,'" rather than just prognostic risk.

The impetus for scores comes from the fact that it is impossible to validate statements about an individual, he said. "It's too small of a group." What can be done, however, is to make statements about people with a given quantitative score and then validate that score. "That's my third main point," he said. "Methods should consider development of action at the individual level, but that action is based on a score, and the score can be validated locally. We can validate whether that score is giving an accurate representation of the expected benefit of being treated as compared with not treated."

In response to a question from the audience, Heagerty acknowledged the importance of knowing and communicating about the methods and data used to generate predictive models for a particular patient. "I feel like we don't do a good job of showing the source data that generated that prediction," he said, "and we could and should." Importantly, researchers should clearly indicate whether the data used to create a model truly applies to the patient. As an example, he referred to an 82-year-old patient discovering that a study used to create a prediction model involved only patients much younger than 82 years—which would be important information. "We have a responsibility to communicate better the evidence base that generates those predictions," he said.

ABSOLUTE RISK VERSUS RELATIVE RISK

In his response to the presentations, Michael Pencina, Vice Dean for Data Science and Information Technology at the Duke Clinical Research Institute raised an issue that would be touched on at various points throughout the workshop: absolute versus relative risk. Absolute risk refers to the chances of something happening over a particular period of time—for example, the chance of a person having a stroke over the coming year. Relative risk refers to the difference in risks between two situations—for example, the risk of having a stroke over the coming year if you take a particular drug versus if you do not. "The absolute risk reduction is a key metric," Pencina said. "It's composed of two

pieces. It's the risk, and it's the relative risk reduction. These two pieces are critical, and you can't focus just on one."

Building on this point, David Kent later commented that "HTE is a scale-dependent concept—you just have to specify the scale. And the issue is that for clinical decision making the most important scale is the absolute scale." The idea that HTE only exists when there is significant variation on the relative scale has served the field poorly, he said. "I think we're a little brainwashed by that distinction. We should always look at the absolute risk scale." Instead of focusing on a search for "statistically significant" relative effect modifiers one variable at a time, the approach needs to provide the kind of evidence that will be helpful to doctors and patients as they are making decisions—one patient at a time.

5

NEXT STEPS FOR IMPLEMENTATION

The development of techniques and models for dealing with heterogeneous treatment effects (HTE) and predicting individual risk is the first step in fulfilling the potential of this field. These techniques and models must then be implemented by clinicians. The next-to-last session of the day was devoted to what it will take to move new capabilities related to HTE into the clinic. Much of the information offered in this session was relevant to the patient question: How can clinicians, as well as the care delivery systems they work in, help me make the best decisions about my health and health care?

Points Highlighted by Individual Speakers

- Models are meaningless without an effective implementation strategy. (Spertus)
- Physician acceptance of a model is critical. (Spertus)
- By integrating prediction tools into a medical records system, it is possible to provide clinicians with near-real-time results and improve decision making related to heterogeneous treatment effects (HTE). (Peterson)
- To improve health care in a world of HTE, it will be crucial to develop better performance measures, specifically measures that take that heterogeneity into account. (Hayward)
- It is important that doctors involve patients in decision making. There are various decision tools that can make this process easier and more effective. (Hayward)

USING HETEROGENEOUS TREATMENT EFFECTS IN ROUTINE CLINICAL CARE

In the area of HTE, almost all effort—and almost all funding—has been focused on determining which patients will do best with which treatments, said John Spertus, Chair and Professor of Medicine at University of Missouri–Kansas City. "What essentially no one is spending any money or research doing," he continued, "is figuring out how we move that knowledge into routine clinical care so that we can start to use it every day on patients to help improve the value of care that we deliver."Yet, that implementation step is equally important, he said. He then described an effort at St. Luke's Mid-America Heart Institute, where he serves as Clinical Director of Outcomes Research, to take that second step and apply knowledge about HTE in helping patients.

For 20 years, the National Cardiovascular Data Registry (NCDR®) has been collecting data on patients and outcomes. "It collects millions of records a year on patients undergoing cardiac procedures, and it builds risk models," he said. After analyzing that myriad data, NCDR provides hospitals with a quarterly report, including the observed versus the expected rates for a variety of complications, among other information. "Twenty years ago, this was sort of state-of-the-art quality assessment through benchmarking," he said. However, Spertus said, there was never much interest in taking those risk models and the data and using them prospectively to improve clinical care—to help doctors and patients make medical decisions that were tailored to the risk of individual patients. In response, he and his colleagues did it themselves, creating a computerized decision-making platform they called ePRISM. He explained,

> The idea was to take the exact risk models that NCDR was using to risk-adjust the performance at hospitals and enter them with patient-specific data and create clinically useful tools that could be part of clinical care that could allow the heterogeneity of benefit for individual patients to be appreciated at the time you're making the decision and treating the patient.

Using risk models for individual patients in this way requires that the models be fully integrated into clinical care "so nobody can get through the process of being treated without getting that risk model run," Spertus said. This process requires a collection of changes to be made in St. Luke's procedures. One important change, he said, was to the hospital's consent forms. They were "terrible," he quipped. The same exact form was used for every procedure, whether it was an angioplasty or a skin biopsy or a liver transplant. It was written at a "16th-grade level," in legalese,

and was exceedingly vague. The form did not educate or inform, and it did not help either the patients or the providers. St. Luke's now uses a personalized consent form generated for each patient. This form is much more readable—written at an 8th-grade level, with no legalese, and with pictures to help explain such things as an angioplasty or a stent. Additionally, one key change, Spertus noted, is that each consent form shows the patient's individualized risk of bleeding or dying from a given procedure, calculated from the risk model as a function of that patient's personal characteristics. This information helps patients make informed decisions, for example, choosing between a bare metal stent versus a drug-eluting stent. For this choice, patients learn that the choice of the bare metal stent makes it more likely that the blood vessel will close within 1 year and require a new procedure, but the drug-eluting stent requires a much longer period of aggressive anti-platelet therapy, which leads to bleeding and bruising and can cause delays in elective procedures. Now, it is up to the informed patient to decide on the trade-off.

For physicians, the ePRISM system provides personalized information about the procedure and the risks on a monitor in the catheterization laboratory (cath lab) as the patient is being seen, Spertus explained. It also provides a personalized recommendation on the approach to be used. "Literally as you are about to touch the patient, you know what [his or her] risks are, and everybody in the cath lab thinks, 'This is how we're going to approach it.'" Spertus and his colleagues tested the system in nine centers around the country with 137 interventional cardiologists, who treated a total 7,408 of patients with a percutaneous coronary intervention (PCI) to insert a stent. As an outcome, they looked at how often bleeding followed the procedure, comparing the rates of post-PCI bleeding in these nine centers before and after the system was put in place. The system brought about a significant reduction in bleeding, Spertus said, and the decrease was greatest for the high-risk patients. "In a fully adjusted model," he said, "there was a 45 percent reduction in bleeding when the doctors knew the risks of their patients before they approached them" (Spertus et al., 2015).

In Figure 5-1, the smooth curve showing the outcome of the trial "hides a lot of messy details," Spertus said, particularly details about the habits of the individual physicians; he then described what he learned about those habits. Ideally, an interventional cardiologist will modify his or her approach to a PCI based on a patient's risk. In particular, for riskier patients, the cardiologist should be using one or more of the well-known bleeding avoidance strategies. But when he examined the records of the 137 interventional cardiologists in the trial, that is not what he found. "This is the scariest research slide I have ever generated in my career," he conveyed. "This is the actual practice pattern of 137 excellent interventional cardiologists at great institutions across the country." (See Figure 5-2.)

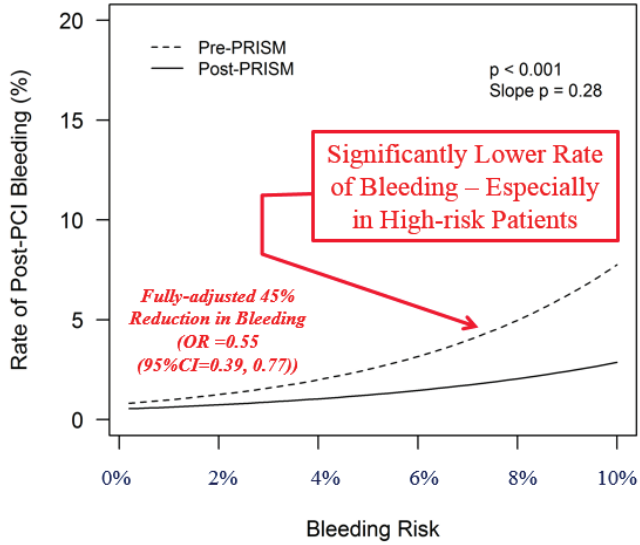


FIGURE 5-1 | Reduction in bleeding after introduction of the ePRISM system.
SOURCES: John Spertus presentation on May 31, 2018; Spertus et al., 2015.

He said,

> What should just be astonishing to you is that it's all over the map. Some doctors are always using bleeding avoidance strategies regardless of risk. That's okay—maybe they're emphasizing safety. However, some doctors are never using them regardless of the patient's risk of bleeding. That makes no sense. And the vast majority of doctors are treating the lower-risk patients more than the higher-risk patients, which is completely counter-intuitive.

A patient who needs an angioplasty and chooses one of these centers is assigned to whichever interventional cardiologist is in the laboratory that day and thus has no way of predicting how his or her risk will be handled. "That's a problem," he said. "We need to be thinking about physician barriers." When he examined the individual performances of the cardiologists in his trial, he found that the physicians tended to fall into one of three categories: those whose performance improved with the use of the system, those who stayed about the same, and those whose performance actually got worse. These results reflected the responses of the doctors to having the new ePRISM system in place. "Some doctors, you give them a risk model, and they're going to improve their performance," Spertus said. "And some doctors … are going to do the exact opposite of what you and your protocol recommend."

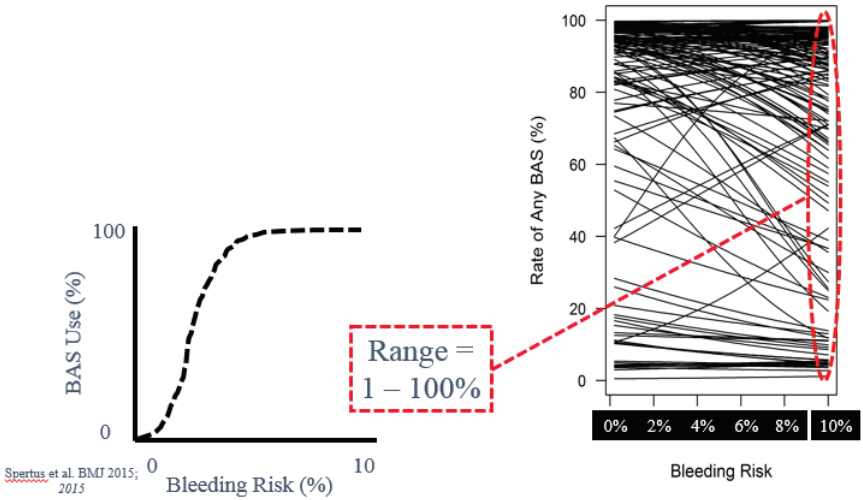


FIGURE 5-2 | The use of bleeding avoidance strategies as a function of bleeding risk in 137 interventional cardiologists.

NOTE: BAS = bleeding avoidance strategies.

SOURCES: John Spertus presentation on May 31, 2018; Spertus et al., 2015.

To understand this situation better, Spertus conducted a study in which he spoke with 27 interventionists at eight centers (Decker et al., 2016). Three themes emerged. The first of which was "experience versus evidence." Some doctors with a lot of experience in the field felt that their own judgment was all they needed. "Some physicians think that they've been doing this for years and years and years and they don't need someone else's tool to help them explain to the patient what they think is important." For these physicians, Spertus said, it will be important to find a way to convince them that the system is not supplanting their experience but rather supplementing it. The second theme was "rationing of care." Some physicians did not like the idea of treating high-risk patients differently from low-risk patients. Spertus quoted one cardiologist who said, "Restenosis is never higher with a drug-eluting stent, never. So … why wouldn't you put the Cadillac in everybody?" But today, with the regular emphasis on reducing health care costs, Spertus said, it is important to push for the use of the technique mainly in those high-risk patients who will most benefit from it—which requires knowing who those patients are. The third theme was the perceived value of the process. Some physicians believed they already knew what the system was telling them, so why did they need a form to tell them what to do? "The point," Spertus said, "is that if physicians alter their behavior and adhere to the risk models, you can

improve the outcomes and the value of health care.… Creating a way for people to embrace the support is very important."

To do that, Spertus and his colleagues developed a five-step program to get doctors to buy in to the process. The five steps are:

- Identifying a clinical champion, a cath lab leader to drive change;
- Creating a risk-based protocol;
- Implementing a standardized timeout;
- Measuring and sharing performance, which provides feedback and accountability; and
- Celebrating success by developing rewards.

When Washington University implemented this program, Spertus said, the post-PCI bleeding rate dropped from 8 to 10 percent down to 2 percent (Spertus presentation on May 31, 2018). "It's such a great reduction that every week one or two patients do not bleed in that cath lab who used to." He noted that about half of the reduction came after the model was implemented, and the rest came after a poster was displayed in the cath lab letting everyone see the percentage of time that each doctor had deviated from the protocol. That feedback cut the bleeding incidence rate in half again.

Toward the end of his presentation, Spertus spoke briefly about shared decision making. According to Spertus, providing patients with a coach to work with when making a decision was a crucial component to improving shared decision making (Ting et al., 2014). When they had a coach, patients were much more likely to have formed a decision on which type of stent they wanted, and they were much more likely to be a part of the decision as to what sort of stent to use, instead of just leaving it to the doctor to decide.

Spertus provided some final thoughts: "Models are meaningless without an effective implementation strategy." Such models must be integrated into the work flow, and simpler models are better. Furthermore, physician acceptance of a model is critical. Compelling evidence of the model's effectiveness is important, but not sufficient. Similarly, proof of its benefit is important, but not sufficient. Incorporating accountability and incentives into the system is critical to its success.

APPLYING PHARMACOGENOMICS IN CLINICAL CARE

The next presenter, Josh Peterson, Associate Professor of Biomedical Informatics and Medicine at the Vanderbilt University Medical Center (VUMC), continued

the theme of dealing with HTE in routine clinical care but spoke in particular about how pharmacogenomics data could be used to make better treatment decisions. To offer some context, he described a 2003 paper (Gandhi et al., 2003) published in *The New England Journal of Medicine* that examined the adverse drug events that happened to patients in ambulatory care. "The bottom line," Peterson said, "is if you give 1,000 patients a prescription . . . and you look at what happens to them in the next 3 months, you'll find that there's a lot of adverse events." Most of these events are mainly an annoyance to the patients, but some are serious. Notably, the 2003 study concluded that 3.8 percent of ambulatory care patients experienced a serious adverse event.

The study focused on preventable events, but a large percentage of the adverse drug events were considered non-preventable, he said.

> Non-preventable ones were things like serious cutaneous adverse reactions, side effects from selective serotonin reuptake inhibitors [SSRIs], nonsteroidal anti-inflammatory drug [NSAID]-related GI [gastrointestinal] events, and beta blocker–related bradycardia, where at that time you shrugged your shoulders and said, "Well, there wasn't much we could do about it."

One exciting aspect of pharmacogenomics, Peterson said, is that it is starting to turn some of these non-preventable events into preventable ones.

Not all adverse events can be avoided yet; there have been, however, some important successes. One such success involves the rare but serious side effect of Stevens–Johnson syndrome in certain patients who are given carbamazepine to prevent seizures. Patients who develop that syndrome get painful rashes and blisters on their skin, Peterson explained, showing photos of patients with severe cases. It was discovered that the syndrome appears mostly in patients with a particular gene variant (i.e., HLA-B*1502) that is common among certain Asian populations. Several Asian countries now require genetic testing before prescribing carbamazepine, Peterson noted, and doing so has dramatically decreased the occurrence of both Stevens–Johnson syndrome and toxic epidermal necrolysis.

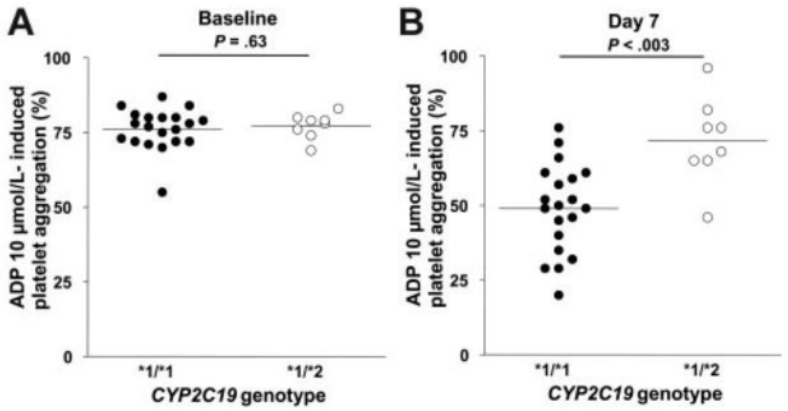
Another critical aspect of such adverse drug events, Peterson said, is that "there is a frustrating lack of drug efficacy for many of the common drugs we use in primary care." About 38 percent of SSRIs are ineffective on average, he said, along with 40 percent of asthma drugs, 43 percent of diabetes drugs, and 50 percent of arthritis drugs. This issue leads to a cycle of "Let's try this drug; well, how about this one?" he said. Such a practice both erodes patients' confidence in the therapies that doctors prescribe and exposes them to a greater number of drugs, each of which has the potential to harm the patient. The goal then, is to learn how to

integrate various factors—age, drug interactions, pharmacogenomics, indication for therapy, behavioral factors, and others—and predict which prescriptions will be most effective for a particular patient and which ones should be avoided; doing so has the potential to provide clinicians with a powerful tool for dealing with HTE.

Before discussing these prediction tools, Peterson spoke about the spectrum of evidence in pharmacogenomics. In particular, he discussed three areas of evidence: (1) analytic validity, or how well a test determines whether a particular gene or genetic variant is present; (2) clinical validity, or how closely the particular genetic variant being analyzed is linked to the manifestation of the disease of interest; and (3) clinical utility, or how helpful the information provided by a particular test will be to a patient. The evidence is strongest for analytic validity. "One of the nice things about genetics," Peterson said, "is you can have a lot of confidence in the fact that you have close to 100 percent reproducibility. If you find a variant, you'll find that same variant, usually with a couple of trouble spots, on multiple platforms."

A vast body of literature concerning clinical validity exists, as well. Genome-wide association studies and phenome-wide association studies are the most common types of studies supporting clinical validity; the results, however, are still too preliminary to form a strong basis for clinical decision making. Peterson listed two other types of clinical validity evidence, correlations with phenotypic testing and candidate gene studies of clinical outcomes. "And this is mostly what we have to work with when we're thinking about implementation," he said. "We'd like to have more clinical utility evidence, which would include randomized trials and comparative effectiveness, but we have to make decisions about what we're going to do now, today, with the evidence that's in front of us," usually without direct evidence of clinical utility.

Peterson then offered specific examples of said evidence. "This is the paper that launched a cottage industry in genetics," he stated, showing a pair of plots from that paper (see Figure 5-3). The paper (Hulot et al., 2006) studied how patients with two variants of the CYP2C19 gene varied in response to the anti-clotting drug clopidogrel. At baseline, the subjects had basically identical platelet aggregation measures (part A of the figure). Yet, 1 week after they were given clopidogrel, those measures looked very different. Patients with the *1/*1 genotype responded to clopidogrel by having about one-third less platelet aggregation, on average, but those with the *1/*2 genotype had, on average, almost no response to the drug. Carriers of the *2 variant also had a poorer clinical response to clopidogrel, Peterson said. Among acute coronary syndrome patients who were undergoing a PCI, the carriers of that variant had significantly more coronary events in the year after they started taking clopidogrel than non-carriers.



Platelet aggregation (absolute values) in response to 10 µM ADP according to the *CYP2C19*2* genotype. Absolute values at baseline (A) and after 6 days (day 7) of clopidogrel 75 mg/d (B). The line indicates the mean value.

FIGURE 5-3 | Platelet aggregation response to clopidogrel varies by CYP2C19 variants.
NOTE: ADP = adenosine diphosphate.
SOURCES: Josh Peterson presentation on May 31, 2018; Hulot et al., 2006.

Finally, Peterson presented data from a randomized controlled trial of azathioprine, a common immunosuppressant used in patients with autoimmune diseases; the effects of azathioprine in the body depend in large part on how various bodily enzymes metabolize the drug (Newman et al., 2011). Thus, as part of the study, the researchers had analyzed enzymatic activity versus the presence or the absence of a particular genetic mutation. "Patients without the variant have a nice bell curve of enzymatic activity," Peterson explained. "Those who are heterozygous for the mutation have their enzymatic activity about half of normal. And the one patient homozygous for the variant has no enzymatic activity at all." The data clearly show how the presence or the absence of genetic variants can shape how individual patients' bodies respond to drug treatment.

With that, Peterson moved on to describe the Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) program being used at VUMC to help doctors shape their prescribing to fit the particular characteristics, including the pharmacogenomic characteristics, of each patient. PREDICT helps target pharmacogenetics testing to patients in whom the information is most likely to be relevant in the near future. The system collects data on patients at the medical center in various ways. For example, patients in the cath lab may receive an order from their physician to have their CYP2C19 status tested. There is also a group of patients who are essentially institutionally funded to be

tested before they need any of the therapies. "The idea is to get the information into the record before it's useful because otherwise there's a lag that becomes a great implementation barrier," Peterson said. "If you get your genetic test a week after your prescription, you've already incurred a week of risk related to that prescription." There are now 15,000 patients in the program, and there are several drugs targeted for genome testing, including clopidogrel, simvastatin, and warfarin. "The whole concept," Peterson said, "is that most of this happens in a semi-automated fashion so that even if you got tested several years ago, you still get some information pushed to you at the right time."

To get the necessary pharmacogenomic information about how individuals with different gene variants respond to various drugs, the program makes heavy use of the Clinical Pharmacogenetics Implementation Consortium (CPIC®) guidelines.[1] The program also engages various clinical experts from the medical center. The result being that clinicians receive a variety of information about how their patients can be expected to respond to various medications. "We sort of automated the results-to-interpretation pipeline to make it as quick as possible," Peterson said. "As soon as the lab signs off on a result, within just a couple of minutes we get an interpretation that shows up in the patient record in a couple of different places so that people can find these results easily." He then provided several examples of the sort of information that the clinicians receive.

In one example, SLCO1B1 gene testing in a patient being considered for lipid lowering therapy revealed the *5/*5 genotype, which leads to decreased transporter function and a high myopathy risk if simvastatin were prescribed. The interpretation of this information, Peterson said, was, "Prescribe a dose of 20 mg or lower, or consider an alternative statin; consider routine CK surveillance."

A second example illustrated the sort of warning a clinician would receive when prescribing clopidogrel to a patient who needed antiplatelet medication but had a CYP2C19 variant that might limit his or her response to clopidogrel. The notice indicates that the patient has a gene variant that is associated with a poor response to clopidogrel and offers some alternatives—in this case, prasugrel or ticagrelor.

The third example was a warfarin adviser who offers a recommended initial dose of warfarin based on a patient's genetic variants and various other factors. "This is actually a pretty popular form of clinical decision support in our institution, meaning that we get an 80 to 90 percent acceptance rate of the offered dose," Peterson said. "The bottom line is, this is how much warfarin we think you should give as a starting dose, and we get a lot of uptake on that." On the other

[1] See https://cpicpgx.org/guidelines (accessed August 9, 2019).

hand, he added, if the clinicians are asked to prescribe a different drug, "that's cognitively a bigger deal for them, they end up accepting that advice, depending on the scenario, 30 to 60 percent of the time."

Finally, he showed a form that is provided to patients informing them of their drug sensitivities. It is made available through the medical center's patient portal. "I don't know if this is completely adequate in terms of educating them about their pharmacogenomic results," Peterson noted, "but it is a very scalable way for patients to get access to results so that if they go outside our institution, they still have a way to refer back to the kinds of genetic variants we found."

IMPROVING PERFORMANCE MEASURES

Rodney Hayward, Professor, Department of Internal Medicine and Department of Health Management and Policy of the University of Michigan and the Ann Arbor Veterans Affairs Medical Center (VAMC) opened his presentation with a point that he would return to several times. "So much of what's wrong with medicine," he said, "is we want to pretend that we can just talk about the one dimension that we care about at a time. Right now, we're thinking of quality, and we can't think of cost when we're talking about quality. We can't think of patient autonomy when we're talking about quality."

To illustrate, he presented a slide with several typical medical targets: hemoglobin A1c less than 7 percent, blood pressure less than 135/90 mm Hg, low-density lipoprotein (LDL) less than 100 mg/dl, and having had an eye exam within the past year. These goals are assumed to be ones that every patient should strive for, Hayward said. "They don't consider the heterogeneity among people, and they don't talk about patient preferences." He then went into detail for each of these typical medical targets.

Since 1997, many have known that the standard target for blood sugar is too stringent, Hayward said. That target is getting A1c—a measure of the average level of blood sugar over the past 2 to 3 months—below 7 percent. "Almost all the benefit [of lowering blood sugar] in people with diabetes is getting them to 8 percent," he said (see Figure 5-4). "But the quality measure is getting them below 7 percent. Most people with type 2 diabetes get almost no benefit from going from 8 to 7 percent. We've known this for [more than] 20 years. It still has not changed."

Something similar is true for the standard 135/90 mm Hg target for blood pressure, Hayward continued. For people with high blood pressure who are at high risk of morbidity and mortality, studies show that the best approach is to prescribe them three or four blood pressure medicines, if tolerated, and not to

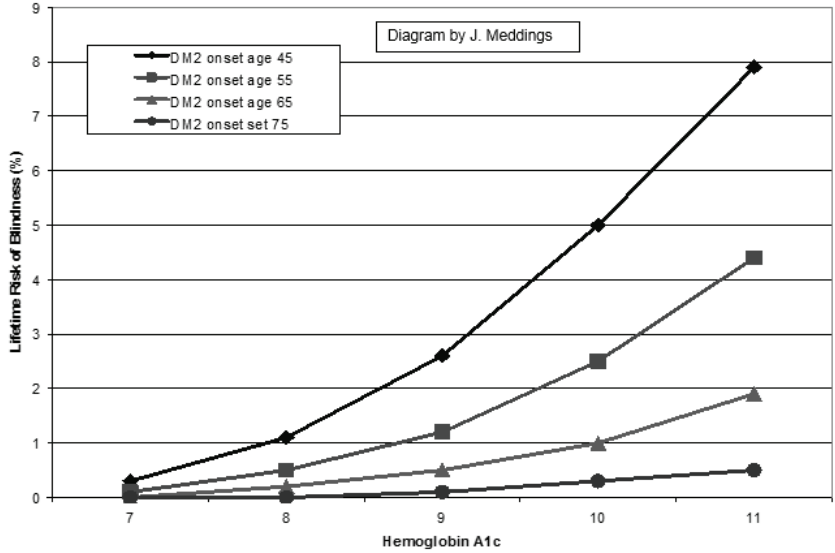


FIGURE 5-4 | Relationship between A1c and lifetime risk of blindness.
NOTE: DM2 = type 2 diabetes mellitus.
SOURCES: Rodney Hayward presentation on May 31, 2018; Vijan et al., 1997.

worry about whether the blood pressure reaches the target (Basu et al., 2016, 2017; Karmali et al., 2018; Sussman et al., 2013; Timbie et al., 2010). Among the high-risk patients, nearly 50 percent will not reach their blood pressure goal, which is not an issue because the use of the medications is most important. With regard to low-risk people with high blood pressure, "once you have them on one or two medicines, you're close to doing net harm."

The situation is the same with lipids, he said. High-risk patients with high levels of LDL benefit from taking medications for lowering LDL levels, but they get little to no additional benefit from the LDL levels dropping; and low-risk patients with high LDL get little benefit from the drugs at all. In short, he said, "Your blood pressure and your LDL do not modify treatment effects much at all, and it's what we base all our guidelines on.… Our dumb quality measures are leading to dumb care."

The issue with eye exams is different, yet the bottom line is the same. Guidelines call for annual eye exams, but few of the problems that eye doctors see are the result of people not having yearly screenings. About two-thirds of these problems, Hayward explained, are due to failures to follow up with patients who have known retinopathy. About one-third is due to people who are seldom or never screened, who have gone years without an eye exam, and less than 1 percent of these

problems are preventable by annual screening, he said. "We are encouraging waste and harm with all of our performance measures because we do not understand heterogeneity of treatment effects," he said.

To improve health care in a world of HTE, it will be crucial to develop better performance measures—measures that take that heterogeneity into account. A good place to start, he said, is with the recognition that health care cannot be dichotomized. The typical approach to performance measures is an either/or affair: either the patient should get the treatment or not, and there is a clear cutoff. "We argue that it is a continuum," he noted. If net value to the patient is plotted along a line, if you are far enough to the right, the benefits to the patient are so great that it is "a no-brainer." Every patient in that position should get the treatment. If you go far to the left, the benefits are questionable and do not outweigh the costs, hence treatment would not be recommended. In the middle, however, is an area of low to moderate net value for which the patient's choices become important—that is, when considerations other than the strictly medical ones come into play.

Several problems also exist with dichotomous performance measures, Hayward said, and he identified two types of these measures, strict and lenient. The strict measures have the following weaknesses, he said:

- They do not target those patients most likely to benefit. More generally, they ignore the heterogeneity of patient risk factors.
- They do not help providers optimize or do the "right thing." They are blunt instruments with little or no clinical nuance.
- They do not take into account patient preferences, and they often mandate care that is not wanted by well-informed patients.
- They can result in unintended consequences, such as wasteful spending.

Lenient dichotomous performance measures have their own problems, he continued:

- They do help target patients who are most likely to benefit. However, they do not promote doing the "right thing." They do not lead doctors to think about optimal care.
- They do not take patient preferences into account. They ignore the treatments with low to moderate net value that may be of interest to patients.
- They, too, can result in unintended consequences. For example, doctors could focus on the high-benefit people and leave a large number of patients behind because the potential benefit for them is not as high.

What are the alternatives? One approach, he suggested, would be to weigh the quality measures by quality-adjusted life years (QALYs) at risk or some other factor. In particular, a performance measure should take into account—and penalize—overtreatment. Performance measures should also consider individual attributes that modify the absolute risk ratio. They should consider effective, safe treatments that have not yet been deployed. And, he added, "You have to consider partial credit. I have people refuse flu shots all the time." Patients will not always do what doctors recommend, and doctors should determine how hard to push according to how important the recommendation is.

It is also important for doctors to involve patients in decision making, Hayward said. To make this process easier and more effective, there are various decision tools that can help explain the situation facing the patient and offer recommended choices. When there is a clear yes or no recommendation, the doctor should inform the patient; but the strength of the recommendation can vary, and the doctor should respect the patient's veto, if it happens. He also recommended that doctors identify the factors that are likely to be most influential in helping the patient decide and present those first.

Next Hayward discussed what he called "finding the preference sensitive zone"—that is, knowing when medical considerations make the answer obvious and when a patient's preferences should come into play. He illustrated this zone with a figure mapping out when a doctor should recommend a patient take aspirin daily to decrease the risk of strokes (see Figure 5-5). In the upper right corner, with high-risk patients who get a large expected increase in QALYs from taking the aspirin, the recommendation is clear: Take aspirin. On the other end, when there is minimum benefit, he would recommend against treatment. And in the middle is where the doctor should talk with the patient about preferences, Hayward said. "In here I say, this is a tough call, but if you don't mind taking an aspirin a day, this could be reasonable because the main thing that would make this not a good idea is not liking to take an aspirin a day. And that's because your risk is low, and your benefit is low."

The preference sensitive zone for lung cancer screening looks very different, however. Risk by itself did not lead to clear decisions; but when risk was combined with life expectancy, it led to very clear recommendations. For people at high risk of developing lung cancer and a life expectancy of more than 10 years, the strong recommendation was always to do screening. "There was no scenario where there was net harm," Hayward said. "You could hate CT [computed tomography] screening. You could double the false positive rate. It was always net benefit." Conversely, for people with limited life expectancy or a low risk of developing lung cancer, the net benefit was low, and the decision was best left

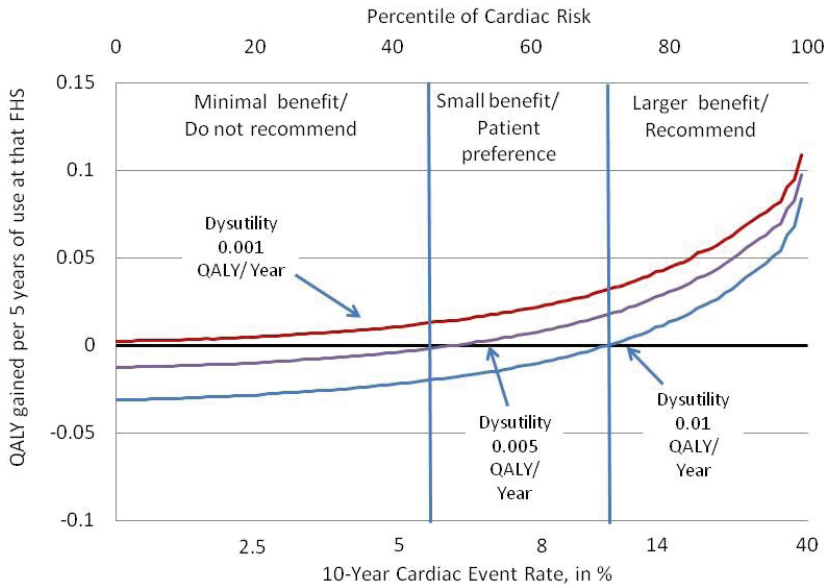


FIGURE 5-5 | Finding the preference sensitive zone.
NOTE: QALY = quality-adjusted life year.
SOURCES: Rodney Hayward presentation on May 31, 2018; Sussman et al., 2011.

to the individual. But the point to keep in mind, Hayward said, is that "this is doable." Such decision support systems can be put into practice.

Finally, he concluded by saying that "we have to have a performance management system" running in parallel with decision support. The "performance overlords" are responsible for making it easy to optimize performance, he said. "It's their responsibility for making these more sensitive, for managing optimal care. Do not let them off the hook. If they give us the tools and they make the measures responsible, then it is our fault. Until then, the over- and under-use is their fault." The point to remember, he said, is that if people focus only on getting the science right and letting clinicians know about the science, "we will not get there. You need accountability and feedback, and you need tools to make it easier."

IDENTIFYING CLINICALLY MEANINGFUL HETEROGENEOUS TREATMENT EFFECTS

"I'm actually very impressed with how much heterogeneity is in every aspect of the system," said Naomi Aronson of the Blue Cross Blue Shield Association, but she added that it is important to keep in mind that not all heterogeneity

is equally important. The classic prototype for heterogeneity of response is the gene-driven response to targeted cancer therapies, for which people with a certain variant will respond well to the treatment and those with a different variant will not respond at all. "It's very directive," she said. "What's important is that it separates patients very clearly." That is a critical question to ask about any heterogeneity, she said: Does it separate? Does it provide sufficient evidence to tell a clinician how different patients should be treated? Another issue is whether the heterogeneity that is visible in a retrospective analysis will actually be apparent prospectively so that it can be used in directing treatment.

It is also important, Aronson said, to distinguish between patient preference and HTE. The ultimate goal is to learn enough about heterogeneity for patients to be able to express their preference and to offer a truly informed consent. "I would urge us to keep these dimensions clarified," she continued, "or it really will confound our purpose in that decision making."

Finally, she warned that developing tests and treatments that take HTE into account may have some unanticipated consequences. If a particular treatment is only effective for a small percentage of people with a disease, and if it becomes possible to identify that small percentage, then the cost of that treatment will be spread over a much smaller group of patients. "Historically, I would say that non-responders have in some way subsidized responders," she said. "Companies can make treatments for a large population, which spreads the cost, so the average cost is less." If it is possible to separate out the responders from the non-responders, the average cost of the treatment will be higher, and, furthermore, the non-responders could end up feeling left out, disenfranchised, and unfairly treated, since they are deselected from receiving therapy (albeit a therapy that is likely to be ineffective for them).

Building on Aronson's comments, Katrina Armstrong, Chair, Department of Medicine, and Physician in Chief of Massachusetts General Hospital said that she had come away from the workshop discussions with four thoughts. First, she said, it will be important to identify criteria for determining which sorts of HTE truly matter. "At the policy level or at the operations level, what I'm really faced with is a ton of potential decisions," she said. "Out of all the decisions that I'm facing, where does heterogeneity really matter?" It will be important to be able to tell the difference between a "no brainer" decision and decisions that will require a lot of time and resources so that these decisions can be appropriately triaged. It will also be important to determine which decisions will require coaches to help patients understand all the different aspects. "I can't hire coaches for everybody," she said. "What metric can I use to say, 'This is a decision where there is a ton of heterogeneity, and there is something I need to pay attention to?'"

Second, she asked, what is it that we are really trying to predict with HTE? "It's not the 5-year trial outcome," she said. "I'm trying to predict my ability to get the patient to the next visit without having hurt them a lot." Thus, it is important to determine what, from a clinician's and patient's perspective, really needs to be predicted. "It is not a single-point decision," she said, "but it is a journey that we are taking with a patient."

Third, she mentioned that much of what predicts how well a patient does in treatment is not found in the clinical variables but rather social factors—whether a patient has housing or insurance or social support. So how can one really understand the HTE when so many of the determinants of that heterogeneity are not clinical variables? "I think it's critical that we try to understand those social variables and how their predictions play out at the same time that we're looking at diving deep into the clinical variables."

Finally, she said, it is important to remember that no data are value-free. As an example, she described how Google Translate takes sentences from Turkish—when the pronouns have no gender—and translates them into English. The Turkish pronoun "o" can mean he, she, or it, depending on the situation. If you translate the Turkish sentence "O [cooks]" into English, she said, Google Translate gives you "She cooks." But if you translate the Turkish sentence "O [operates]," it comes out as "He operates." Google Translate uses its vast database to guess whether a masculine or feminine pronoun is the more likely choice for a particular situation and uses that to settle on "he" or "she." It is just data behind the decision, but that decision is not a value-free one.

DISCUSSION

The first question during the discussion period concerned where to start with convincing doctors to use HTE in their practices. Sheldon Greenfield noted that, judging from some of the workshop presentations, generalists seemed less resistant to adopting HTE recommendations than specialists such as cardiologists. Spertus responded that the most important thing is just to start somewhere. Target a few areas, develop the necessary tools, and build the necessary culture, he said, but do it now. Do not delay.

Hayward had a different answer. It will not work to ignore the specialty areas, he said. "You need to deal with the sub-specialty societies," he explained. "There is no other way. If you go around them, they will win the political battle. All it takes is one prominent cardiologist from Harvard saying, 'You're killing people.'" Instead, he said, it is important to find people who are prominent and connected

but open-minded and enlist them to be on your side. You need to be willing to spend the time working to convince such people.

The conversation then expanded to the more general question of what sort of approach it will take to get HTE widely adopted by the health care community. Hayward said that the health care community must recognize that it will inevitably be a long process requiring both a long-term strategy and patience. He quoted Bill Gates as saying, "People dramatically overestimate how much change they can make in 1 year and underestimate how much change they can do in 10."

Spertus noted that some changes in medicine do happen quickly. It is not always clear why something is adapted so quickly, he said, but psychological factors clearly play a role. Another approach would be to find economic incentives because that will get the attention of the institutions. "There is an opportunity for a lot of creative thinking about creating the incentives to accelerate the change that we're talking about," he said. "We just have not been doing enough of that creative thinking to figure those out yet."

6

A RESEARCH AGENDA FOR PERSONALIZING CARE AND IMPROVING TREATMENT OUTCOMES

The last session of the day was devoted to a look to the future. Workshop participants discussed what is needed to reach the point when the methods for understanding heterogeneous treatment effects (HTE) are more fully developed, as are the tools and the approaches for translating the findings to inform decisions at the point of care (see Box 6-1). The discussion drew on points raised throughout the day to develop a research agenda for the field moving forward. Making progress will require a research agenda focused not only on improving the methods for discovering HTE, but also an agenda focused on best practices for implementing risk models at the point of care and on payment policies that support the effective targeting of treatments.

DESIGNING RESEARCH TO MEET THE NEEDS OF END-USERS

Joseph Selby commented that supporting research designed to understand HTE is extremely relevant to the Patient-Centered Outcomes Research Institute's (PCORI's) mission, as it is a central component of patient-centered outcomes research.

Selby then identified four future directions that emerged from the workshop:

- Improve the quality and the availability of clinical data, as well as data from clinical trials, so they can be used to understand HTE.
- Reform the clinical research process, and particularly the pre-approval research process, so trials are designed to understand HTE and "we're not hit by new therapies for which there is no evidence to help guide who would actually benefit from them."

BOX 6-1

Summary of Priorities That Participants Identified as Appropriate for Research on Predictive Approaches to Heterogeneous Treatment Effects (HTE)

- Better understand the value of these methods through empirical analyses across a wider range of clinical domains.
- Identify heuristics or general principles to judge the adequacy of sample sizes for predictive analytical approaches to HTE.
- Determine optimal approaches to methods that permit the exploration of relative effect modifiers while strongly protecting against false positive findings.
- Better operationalize an approach to evaluating the a priori credibility of relative effect modifiers for inclusion in treatment effect models.
- Determine the optimal measures to evaluate models intended to predict treatment benefit.
- Better understand the impact of different missingness mechanisms and develop principled methods for dealing with missing data in the context of subgroup identification.
- Determine methods to permit models predicting treatment effect to cope with missing data in clinical practice.
- Develop a better understanding of data-driven approaches to predicting patient benefit, including machine learning techniques.
- Determine optimal methods to achieve balance in covariates across subgroups in observational data to reliably measure HTE.

- Determine the appropriate role for observational data in research on HTE, as "there's probably a big role for the large observational data that many can now muster."
- Understand how to implement the findings of research on HTE, so they are used to inform shared decision making. Ultimately, Selby said, shared decision making will likely be more important in dealing with HTE than coverage decisions by insurance companies.

Related to these future directions, Evelyn Whitlock, Chief Science Officer at PCORI, offered a framework for what will be required to move an HTE research agenda forward (see Figure 6-1). "This is a figure that we developed for the work

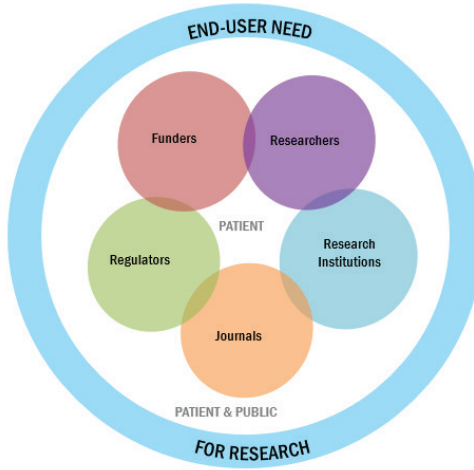


FIGURE 6-1 | Levers for improvement in the research ecosystem.
SOURCE: Evelyn Whitlock presentation on May 31, 2018.

that we're doing internationally with other research funders looking at what are the levers for improvement to reduce waste and improve value in the research ecosystem," she explained.

"As many of you are aware," Whitlock continued, "there has been a movement internationally to look at avoidable waste in research investment," as well as a variety of other factors that are important to moving the agenda in this area forward, including starting with the end-user in mind and making sure that all research results are available and that the associated data can be accessed by other scientists. As Whitlock explained it, the basic idea underlying the framework is that research should be focused on meeting end-user needs. Building on earlier points made by Thomas Concannon, Seth Morgan, and Christine Stake, Whitlock reiterated that "we need to start with the end in mind, we need to know what's going to help patients. We need to do it with the involvement of patients and the public."

At present, she said, PCORI is working to decide on sensible next steps for this research in the coming years. "This is a meeting that illustrates the commitment that PCORI has made in this area." Specifically, she continued, one of PCORI's goals is to assemble basic methods for understanding HTE, particularly tools for outcome risk prediction. "If you can accurately predict outcome risk," she said, "then even if you don't have treatment effect modification, you're going to have more benefit in the higher-risk people." Thus, PCORI is interested in determining what evidence is needed to move forward with these various tools—and also in figuring out if perhaps there are areas for which the evidence is

already sufficient. "Do we have a cadre of established, validated prognostic models that could come off the shelf for some of these situations?" she asked. "There may be more than we're aware of."

A RESEARCH AGENDA FOR UNDERSTANDING AND LEVERAGING TREATMENT HETEROGENEITY TO IMPROVE PATIENT CARE

Steven Goodman, Associate Dean for Clinical and Translational Research at Stanford University, who is a co-chair of PCORI's methodology committee, provided a thoughtful discussion of a number of the philosophical and methodological issues that will need to be grappled with if the application of HTE is to reach its full potential.

One of his main research interests, Goodman said, is the foundations of scientific and biomedical inference or, as he put it, "How do we know that the things we saw are true?" That, he said, was the "fundamental dilemma" underlying much of the discussion that had taken place at the workshop. There are foundational issues facing the field "that we cannot get around," he said. One of those issues concerns causality and how one determines it. According to Goodman, "Whether one phenomenon that's predicted by another phenomenon is causal is not found in the data. So, we have to bring other things to the data to determine causality."

Another key issue is the nature of risk and probability, he said. What is a risk? "It is perhaps the only biomedical property that we cannot measure in the individual." One can measure such things as height and weight directly from an individual, but determining risk requires working with a group, he noted. You measure risk for that group and then assign the group risk to the individual members of the group. That in itself is a huge leap philosophically, Goodman said—to assume that the risk of the group is the risk to the individual—but having made that leap, one is then faced with a crucial question: What is the right group? This is the reference class problem that David Kent described, Goodman noted. "It turns out," he said, "that the right group is the group defined by the causal factors of the phenomenon that you're studying." But what are the causes of the phenomenon? That's what you were trying to find out in the first place. "And now we're in a circle," he said. "This is an irreducible dilemma. We will always be faced with this dilemma, and many of the debates that we had here today are just transmuting this dilemma into other questions."

From those two rather philosophical issues, Goodman transitioned into some practical concerns surrounding the study of HTE. The first related to the issue of the likely proliferation of risk-prediction tools as HTE is incorporated into an increasing number of randomized controlled trials (RCTs). "One of the worries

in the decision to start developing risk-stratification or risk-prediction tools in every RCT is that what we will have is a proliferation of these risk predictions based on every RCT," he said. This is why it is critical to have standard risk-prediction tools, he continued, but very few of those have been developed. Even after a number of risk-prediction models have been developed, it will be a huge challenge to come to agreement on which are the best to use. Noting that it had been suggested during the workshop that risk and benefit models should be developed for every trial, Goodman cautioned, "I think we're going to have to be very, very careful about how we do that."

Furthermore, to the extent that the models are predictive and not just prognostic, the issues become even more complex, Goodman said, "because then we're getting into the issues of causality." RCTs were developed to assess causal effects, and moving away from the standard RCT model will offer challenges. "I think there are a lot of benefits to come," he said, "but we're going to have to be very, very, very careful as we migrate from causal inferences based on randomization to causal inferences based on models.... And I think a number of people have pointed that out."

A related issue is research reproducibility. As Sanjay Basu demonstrated, two studies of the same treatments can show opposite effects, with one demonstrating net benefit and the other demonstrating no net benefit or net harm. Some of the variation is a result of the eligibility criteria. Those eligibility criteria are an initial reference class, a first guess at which group is likely to benefit from the treatment. "If we start deviating from that reference class and say that only certain ones of these are going to benefit," he continued, "then we have this question of, Should we, or do we, only focus future RCTs on that subgroup? Or, do we do the reverse? Do we expand the eligibility criteria for the RCT because we want to get information on treatment benefit for everybody?"

Goodman said he felt that there is a tension in the field concerning whether to restrict treatment or expand treatment. In the workshop, he said, he heard arguments both ways, with some saying it should be expanded and others saying it should be restricted. Once a treatment becomes widely used, a related question arises of how to decide which subgroups get the treatment paid for. Goodman urged thinking about the question in terms of the collective population benefit. Sometimes the most population benefit may come from treating the 10 percent at the highest risk, while at other times the greatest benefit may arise from treating the other 90 percent. It is likely that the answer will be different depending on the treatment and the condition, Goodman said. "Where and how we set that cutpoint might be an issue of politics, it might be an issue of economics, but it's not a given mathematically where that trade-off needs to be."

Next, he addressed a comment by Frank Harrell that HTE should not be used to rescue failed trials. "I would say, Why not?" Goodman said. If people use HTE tools to examine successful trials and identify subgroups for which the treatment does not work, why not examine trials that show moderate effects that are not statistically significant and look for subgroups of patients who actually benefit from the treatment? "There may be resource reasons why we don't want to do that," he said, but "I'm worried about saying that we can't use it just from a logical standpoint." Harrell asked for further discussion of this issue.

Then, referring to comments by Naomi Aronson, Katrina Armstrong, and Rodney Hayward, Goodman said that it will be important to think about what sorts of social factors should be incorporated into models. Social factors can influence personal preferences, compliance, and other factors that can play a role in a model's calculations. Just how much the models should incorporate remains an open question, he said.

Another major question regarding the models is how to determine their effectiveness. The best option he sees is to use RCTs to test the models, just as RCTs are used to determine the effectiveness of diagnostic tests. In one arm, for example, patients would be treated according to the results of a risk-stratification model—which might mean that some patients do not get treated at all—while on the other arm, patients are treated the traditional way, without a model to guide the treatment. As far as judging the quality of evidence from the various models being developed to deal with HTE, "I don't know that we're even at the beginning," he said. "So how are we going to grade recommendations on treating high-risk patients or treating patients with a particular multi-factorial risk–benefit profile from these models? I have barely a clue." But it is important to start thinking about it now, he said, "because if we cannot figure out what the reliability of this evidence is, we will be caught on the same horns of the dilemma that the guideline developers were in the 70s and 80s when we first started to learn about relying on observational data and clinical trials of varying quality."

Finally, Goodman said that given what John Spertus had said about the difficulty of getting clinicians to follow even very simple rules from RCTs and systematic reviews, "I worry a lot about the prospect of implementation for these far more complex guidelines." In conclusion, Goodman said, "We need a research agenda on these models, a practice and implementation agenda…. We need a payer agenda to figure out whether the use of these things should guide what is reimbursed. We need a patient decision-making agenda, and I think we need a political agenda because this is a different paradigm." The precision medicine paradigm has already broken the ice and prepared the way for the HTE paradigm, Goodman said, "but I think 95 percent of that [precision medicine] is hype. So,

we have to be careful that we focus on the meat here and that we actually use these in a way that does more good than harm."

DISCUSSION

During the discussion period, Robert Temple from the U.S. Food and Drug Administration (FDA) raised the issue of why so many clinical trials tend to have people who are very sick—a choice that can make it harder to observe HTE and to determine the net benefits of the treatment for those who are less sick. The reason, he said, is that it allows the researchers to get more "hits" and to test the effectiveness of the treatment for less money than it would take if the subjects were less sick. So, he said, "in cardiology the first study we get is in people who are high risk. If you want to know if the drug works in anybody, that's how you find out." It is called "prognostic enrichment," he said.

That triggered a wide-ranging discussion of inclusion and exclusion criteria for clinical trials. One alternative, Ravi Varadhan, Associate Professor of Oncology at the Johns Hopkins Center on Aging and Health, commented, would be to choose subjects for trials in a way that is parallel to how surveys are done, with careful attention paid to obtaining a representative sample of the population. Jesse Berlin, Vice President and Global Head of Epidemiology from Johnson & Johnson, suggested that "there ought to be a way to build randomization in a pragmatic way into actual clinical practice" so that the results of clinical practice could be used in the same way as RCT data. There would be ethical issues to be discussed, he acknowledged, but "the idea is to turn this into a real learning health care system."

Sheldon Greenfield suggested combining RCTs with observational studies. The Women's Health Study did something similar, he said. Steve Goodman agreed that combining observational studies and RCTs was important, as is combining analyses from multiple observational studies, especially since "there are a lot of initiatives going on right now to mimic RCT evidence with appropriately designed observational evidence." There are many domains for which observational evidence is very important, he said, and others for which it is not.

Robert Golub, a Deputy Editor of the *Journal of the American Medical Association* (*JAMA*), spoke about issues related to communicating HTE results. "I would like you to think about how to communicate these types of findings within journal articles to clinician readers," he said. "I am convinced that most of our readers do not really understand most of the things that *JAMA* publishes. They may understand the basic outlines of an RCT, but that is pretty much

it." Communicating HTE results accurately will be even more difficult than communicating about RCTs, he said, so it is important for those in the field to identify effective ways to communicate the concepts and help clinicians understand the nuances of the work. It is not enough just to provide tools to tell clinicians what treatments to prescribe in which situations, he said—that is just turning clinicians into technicians. "Clinicians need to understand the research that is behind that."

CONCLUSIONS

As stated by Whitlock, numerous speakers over the course of this workshop provided convincing demonstrations that variations in baseline outcome risk can be expected to influence absolute treatment effects in treatment-eligible patients, that meaningful variation in outcome risk is quite common among trial participants and treatment-eligible populations, and that the subset—and often a minority—of trial participants who are at higher baseline risk for the outcomes the treatment addresses will often drive the finding of overall benefit.

However, while there are examples of risk models being used to tailor care, the methods for modeling these effects and for implementing those models in clinical care to personalize treatment decisions are still in their infancy. In order to facilitate progress, the field must not only address outstanding methodological questions, it must also determine best practices for implementing risk models and predictions tools in clinical practice so they can be used by patients and clinicians at the point of care to inform treatment decisions and consider appropriate value-based payment models that effectively target treatments to subpopulations that are most likely to benefit. Therefore, key directions for the field include

- Developing guidance on approaches for assessing the effectiveness or validity of predictive and prognostic models;
- Understanding the comparative performance of supervised machine learning methods that can be applied to understand HTE;
- Facilitating collaboration and leadership across various sectors of the research ecosystem to create prioritized opportunities for large trial re-analyses or collaborative individual patient data analyses to examine the HTE most likely to impact population health;
- Describing approaches to implementing risk models in clinical care and providing guidance on which approaches are most effective at informing decisions both at the point of care and at the level of the health care system;

- Considering approaches for integrating data related to the social determinants of health into risk prediction models;
- Determining the role for observational data and when it is appropriate to combine RCTs and observational data;
- Reforming the predominant fee-for-service payment system in the United States to one that rewards value and population health improvements;
- Promoting dissemination of innovative trial designs, including those sampling larger and broader populations to enrich patient heterogeneity; and
- Establishing or extending research reporting guidelines to promote the conduct of predictive HTE analyses.

Understanding HTE can transform medical care by increasing the likelihood that patients will benefit from the treatments that are offered to them and by contributing to the goal of avoiding harmful or wasteful treatment choices. Patients want precise answers about how a given treatment is likely to work for them, given their unique individual characteristics. A one-size-fits-all approach to treating a medical condition based on average responses from clinical trials is inadequate; instead, treatments should be tailored to individuals based on heterogeneity of their clinical characteristics and their personal preferences.

REFERENCES

ACCORD Study Group, W. C. Cushman, G. W. Evans, R. P. Byington, D. C. Goff, Jr., R. H. Grimm, Jr., J. A. Cutler, D. G. Simons-Morton, J. N. Basile, M. A. Corson, J. L. Probstfield, L. Katz, K. A. Peterson, W. T. Friedewald, J. B. Buse, J. T. Bigger, H. C. Gerstein, and F. Ismail-Beigi. 2010. Effects of intensive blood pressure control in type 2 diabetes mellitus. *The New England Journal of Medicine* 362(17):1575–1585.

Athey, S., and G. Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Basu, S., J. S. Yudkin, J. B. Sussman, C. Millett, and R. A. Hayward. 2016. Alternative strategies to achieve cardiovascular mortality goals in China and India: A microsimulation of target-versus risk-based blood pressure treatment. *Circulation* 133(9):840–848.

Basu, S., J. B. Sussman, J. Rigdon, L. Steimle, B. T. Denton, and R. Hayward. 2017. Benefit and harm of intensive blood pressure treatment: Derivation and validation of risk models using data from SPRINT and ACCORD trials. *PLoS Medicine* 14(10):e1002410.

Calfee, C. S., K. Delucchi, P. E. Parsons, B. T. Thompson, L. B. Ware, M. A. Matthay, and the NHLBI ARDS Network. 2014. Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *Lancet Respiratory Medicine* 2(8):611–620.

Carey, L. A., and E. P. Winer. 2016. I-SPY 2—Toward more rapid progress in breast cancer treatment. *The New England Journal of Medicine* 375(1):83–84.

Carson, P., S. Ziesche, G. Johnson, and J. N. Cohn. 1999. Racial differences in response to therapy for heart failure: Analysis of the vasodilator-heart failure trials. *Journal of Cardiac Failure* 5(3):178–187.

Chassin, M. R., J. Kosecoff, D. H. Solomon, and R. H. Brook. 1987. How coronary angiography is used. Clinical determinants of appropriateness. *JAMA* 258(18):2543–2547.

Cohn, J. N., D. G. Archibald, S. Ziesche, J. A. Franciosa, W. E. Harston, F. E. Tristani, W. B. Dunkman, W. Jacobs, G. S. Francis, K. H. Flohr, S. Goldman, F. R. Cobb, P. M. Shah, R. Saunders, R. D. Fletcher, H. S. Loeb, V. C. Hughes, and B. Baker. 1986. Effect of vasodilator therapy on mortality in chronic congestive heart failure Results of a Veterans Administration Cooperative Study. *New England Journal of Medicine* 314(24):1547–1552.

Cohn, J. N., G. Johnson, S. Ziesche, F. Cobb, G. S. Francis, F. Tristani, R. Smith, W. B. Dunkman, H. Loeb, M. Wong, G. Bhat, S. Goldman, R. D. Fletcher, J. Doherty, C. V. Hughes, P. Carson, G. Cintron, R. Shabetai, and C. Haakenson. 1991. A comparison of enalapril with hydralazin-isosorbide dinitrate in the treatment of chronic congestive heart failure. *The New England Journal of Medicine* 325(5):303–310.

Decker, C., L. Garavalia, B. Garavalia, E. Gialde, R. W. Yeh, J. Spertus, and A. K. Chhatriwalla. 2016. Understanding physician-level barriers to the use of individualized risk estimates in percutaneous coronary intervention. *American Heart Journal* 178:190–197. https://doi.org/10.1016/j.ahj.2016.03.027.

Frangakis, C. E., and D. B. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58(1):21–29.

Friedman, R. A. 2015. To treat depression, drugs or therapy? *The New York Times*, January 8.

Gandhi, T. K., S. N. Weingart, J. Borus, A. C. Seger, J. Peterson, E. Burdick, D. L. Seger, K. Shu, F. Federico, L. L. Leape, and D. W. Bates. 2003. Adverse drug events in ambulatory care. *The New England Journal of Medicine* 348:1556–1564.

Goldstein, D. B. 2009. Common genetic variation and human traits. *The New England Journal of Medicine* 360(17):1696–1698.

Holtzman, N., and T. Marteau. 2000. Will genetics revolutionize medicine? *The New England Journal of Medicine* 343(2):141–144.

Hulot, J. S., A. Bura, E. Villard, M. Azizi, V. Remones, C. Goyenvalle, M. Aiach, P. Lechat, and P. Gaussem. 2006. Cytochrome P450 2C19 loss-of-function polymorphism is a major determinant of clopidogrel responsiveness in healthy subjects. *Blood* 108(7):2244–2247. doi: 10.1182/blood-2006-04-013052.

Ioannidis, J. P. 2009. Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. *Circulation: Cardiovascular Genetics* 2(1):7–15.

Janes, H., M. S. Pepe, L. S. McShane, D. J. Sargeant, and P. J. Heagerty. 2015. The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of the National Cancer Institute* 107(8):pii:djv157.

Janssens, A. C., and C. M. van Duijn. 2008. Genome-based prediction of common diseases: Advances and prospects. *Human Molecular Genetics* 17(R2):R166–R173.

Janssens, A. C., M. C. Pardo, E. W. Steyerberg, and C. M. van Duijn. 2004. Revisiting the clinical validity of multiplex genetic testing in complex diseases. *American Journal of Human Genetics* 74(3):585–589.

Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under uncertainty: Heuristics and biases. First edition*. New York: Cambridge University Press.

Karmali, K. N., D. M. Lloyd-Jones, J. van der Leeuw, D. C. Goff, Jr., S. Yusuf, A. Zanchetti, P. Glasziou, R. Jackson, M. Woodward, A. Rodgers, B. C. Neal, E. Berge, K. Teo, B. R. Davis, J. Chalmers, C. Pepine, K. Rahimi, and J. Sundström. 2018. Blood pressure-lowering treatment strategies based on cardiovascular risk versus blood pressure: A meta-analysis of individual participant data. *PLoS Medicine* 15(3):e1002538. doi: 10.1371/journal.pmed.1002538.

Kent, D. M., and R. D. Hayward. 2007. Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *JAMA* 298(10):1209–1212.

Kent, D. M., C. H. Schmid, J. Lau, and H. P. Selker. 2002. Is primary angioplasty for some as good as primary angioplasty for all? Modeling across trials and individual patients. *Journal of General Internal Medicine* 17:887–894.

Kent, D. M., P. M. Rothwell, J. P. Ioannidis, D. G. Altman, and R. A. Hayward. 2010. Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials* 11:85.

Khera, A. V., M. Chaffin, K. Aragam, C. A. Emdin, D. Klarin, M. Haas, C. Roselli, P. Natarajan, and S. Kathiresan. 2018. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. *Nature Genetics* 50:1219–1224.

Knaus, W. A., D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, and F. E. Harrell, Jr. 1991. The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 100(6):1619–1636.

Kozminski, M. A., J. T. Wei, J. Nelson, and D. M. Kent. 2015. Baseline characteristics predict risk of progression and response to combined medical therapy for benign prostatic hyperplasia (BPH). *BJU International* 115(2):308–316.

Kravitz, R. L., N. Duan, and J. Braslow. 2004. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly* 82(4):661–687.

Lango, H., the UK Type 2 Diabetes Genetics Consortium, C. Palmer, A. Morris, E. Ziggini, A. Hattersley, M. McCarthy, T. Frayling, and M. Weedon. 2008. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* 57(11):3129–3135.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753.

McGrath, C. L., M. E. Kelley, P. E. Holtzheimer III, B. W. Dunlop, W. E. Craighead, A. Franco, C. Craddock, and H. Mayberg. 2013. Toward a neuroimaging treatment selection for major depressive disorder. *JAMA Psychiatry* 70(8):821–829.

Meehl, P. E. 2013. *Clinical versus statistical precision: A theoretical analysis and a review of the evidence*. Brattleboro, VT: Echo Point Books & Media.

Newman, W. G., K. Payne, K. Tricker, S. Roberts, R. Elltiott, E. Fargher, S. Pushpakom, J. E. Alder, G. P. Sidgwick, D. Payne, R. A. Elliott, M. Heise, R. Elles, S. C. Ramsden, J. Andrews, J. B. Houston, F. Qasim, J. Shaffer, C. E. M. Griffiths, D. W. Ray, I. Bruce, W. E. R. Ollier, and the TARGET study recruitment team. 2011. A pragmatic randomized controlled trial of thiopurine methyltransferase genotyping prior to azathioprine treatment: The TARGET study. *Pharmacogenomics* 12(6):815–826.

Park, J. W., M. C. Liu, D. Yee, C. Yau, L. J. van'tVeer, Fr. Symmans, M. Paoloni, J. Perlmutter, N. M. Hylton, M. Hogarth, A. DeMichele, M. B. Buxton, A. J. Chien, A. M. Wallace, J. C. Boughey, T. C. Haddad, S. Y. Chui, K. A. Kemmer, H. G. Kaplan, C. Isaacs, R. Nanda, D. Tripathy, K. S. Albain, K. K. Edmiston, A. D. Elias, D. W. Northfelt, L. Pusztai, S. L. Moulder, J. E. Lang, R. K. Viscusi, D. M. Euhus, B. B. Haley, Q. J. Khan, W. C. Wood, M. Melisko, R. Schwab, T. Helsten, J. Lyandres, S. E. Davis, G. L. Hirst, A. Sanil, L. J. Esserman, and D. A. Berry for the I-SPY 2 Investigators. 2016. Adaptive randomization of neratinib in early breast cancer. *The New England Journal of Medicine* 375:11–22.

Pharoah, P. D., A. Antoniou, M. Bobrow, R. L. Zimmern, D. F. Easton, and B. A. Ponder. 2002. Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* 31(1):33–36.

Qian, M., and S. A. Murphy. 2011. Performance guarantees for individualized treatment rules. *Annals of Statistics* 39(2):1180–1210.

Rothwell, P. M., Z. Mehta, S. C. Howard, S. A. Gutnikov, and C. P. Warlow. 2005. Treating individuals 3: From subgroups to individuals: General principles and the example of carotid endarterectomy. *Lancet* 365(9455):256–265.

Rugo, H. S., O. I. Olopade, A. DeMichele, C. Yau, L. J. van't Veer, M. B. Buxton, M. Hogarth, N. M. Hylton, M. Paoloni, J. Perlmutter, W. F. Symmans, D. Yee, A. J. Chien, A. M. Wallace, H. G. Kaplan, J. C. Boughey, T. C. Haddad, K. S. Albain, M. C. Liu, C. Isaacs, Q. J. Khan, J. E. Lang, R. K. Viscusi, L. Pusztai, S. L.

Moulder, S. Y. Chui, K. A. Kemmer, A. D. Elias, K. K. Edmiston, D. M. Euhus, B. B. Haley, R. Nanda, D. W. Northfelt, D. Tripathy, W. C. Wood, C. Ewing, R. Schwab, J. Lyandres, S. E. Davis, G. L. Hirst, A. Sanil, D. A. Berry, and L. J. Esserman. 2016. Adaptive randomization of veliparib-carboplatin treatment in breast cancer. *The New England Journal of Medicine* 375(1):23–34.

Simon, R. 2015. Sensitivity, specificity, PPV, NPV for predictive biomarkers. *Journal of the National Cancer Institute* 107(8):pii:djv153.

Spertus, J., C. Decker, E. Gialde, P. Jones, E. McNulty, R. Bach, and A. Chhatriwalla. 2015. Precision medicine to improve use of bleeding avoidance strategies and reduce bleeding in patients undergoing percutaneous coronary intervention: Prospective cohort study before and after implementation of personalized bleeding risks. *The BMJ* 350:h1302.

SPRINT Research Group, J. T. Wright, Jr, J. D. Williamson, P. K. Whelton, J. K. Snyder, K. M. Sink, M. V. Rocco, D. M. Reboussin, M. Rahman, S. Oparil, C. E. Lewis, P. L. Kimmel, K. C. Johnson, D. C. Goff, Jr, L. J. Fine, J. A. Cutler, W. C. Cushman, A. K. Cheung, and W. T. Ambrosius. 2015. A randomized trial of intensive versus standard blood pressure control. *The New England Journal of Medicine* 373(22):2103–2116.

Steyerberg, E. W., A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. 2010. Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology* 21(1):128–138.

Steyerberg, E. W., T. van der Ploeg, and B. Van Calster. 2014. Risk prediction with machine learning and regression methods. *Biometrical Journal* 56(4):601–606. doi: 10.1002/bimj.201300297.

Sussman, J. B., S. Vijan, H. Choi, and R. A. Hayward. 2011. Individual and population benefits of daily aspirin therapy: A proposal for personalizing national guidelines. *Circulation: Cardiovascular Quality and Outcomes* 4(3):268–275.

Sussman, J. B., S. Vijan, and R. A. Hayward. 2013. Using benefit-based tailored treatment to improve the use of antihypertensive medications. *Circulation* 128(21):2309–2317.

Sussman, J. B., D. M. Kent, J. P. Nelson, and R. A. Hayward. 2015. Improving diabetes prevention with benefit based tailored treatment: Risk-based reanalysis of Diabetes Prevention Program. *The BMJ* 350:h454.

Thune, J. J., D. Hoefsten, M. Lindholm, L. Mortensen, H. Andersen, T. Nielsen, L. Kober, and H. Kelbaek. 2005. Simple risk stratification at admission to identify patients with reduced mortality from primary angioplasty. *Circulation* 112:2017–2021.

Timbie, J. W., R. A. Hayward, and S. Vijan. 2010. Diminishing efficacy of combination therapy, response-heterogeneity, and treatment intolerance limit the attainability of tight risk factor control in patients with diabetes. *Health Services Research* 45(2):437–456.

Ting, H. H., J. P. Brito, and V. M. Montori. 2014. Shared decision making: Science and action. *Circulation: Cardiovascular Quality and Outcomes* 7:323-327. https://doi.org/10.1161/CIRCOUTCOMES.113.000288.

Upshaw, J. N., D. van Klaveren, M. A. Konstam, and D. M. Kent. 2018. Digoxin benefit varies by risk of heart failure hospitalization: Applying the Tufts MC HF risk model. *American Journal of Medicine* 131(6):676–683.

Vijan, S., T. Hofer, and R. A. Hayward. 1997. Estimated benefits of glycemic control in microvascular complications in type 2 diabetes. *Annals of Internal Medicine* 127(9):788–795.

Wallentin, L., R. C. Becker, A. Budaj, C. P. Cannon, H. Emanuelsson, C. Held, J. Horrow, S. Husted, S. James, H. Katus, K. W. Mahaffey, B. M. Scirica, A. Skene, P. G. Steg, R. F. Storey, and R. A. Harrington for the PLATO Investigators. 2009. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *The New England Journal of Medicine* 361:1045–1057. doi: 10.1056/NEJMoa0904327.

Wang, J., M. R. Ban, G. Y. Zou, H. Cao, T. Lin, B. A. Kennedy, S. Anand, S. Yusuf, M. W. Huff, R. L. Pollex, and R. A. Hegele. 2008. Polygenic determinants of severe hypertriglyceridemia. *Human Molecular Genetics* 17(18):2894–2899.

Wennberg, J., and A. Gittelsohn. 1973. Small area variations in health care delivery. *Science* 182(4117):1102–1108.

Yang, Q., M. J. Khoury, L. Botto, J. M. Friedman, and W. D. Flanders. 2003. Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *The American Journal of Human Genetics* 72:636–649.

Zimmer, C. 2018. Genetic intelligence tests are next to worthless. *The Atlantic*, May 29. https://www.theatlantic.com/science/archive/2018/05/genetic-intelligence-tests-are-next-to-worthless/561392 (accessed July 2, 2018).

Appendix A

GLOSSARY

Area under the receiver operating characteristic (ROC) curve (AUC): A measure of the discrimination of a logistic regression model. The ROC curve is the plot of sensitivity versus one minus specificity over all possible thresholds of predicted probability. The area under the ROC curve is numerically equivalent to the c-statistic for a binary outcome.

Bayesian nonparametric methods: An approach to model selection that allows the data to determine the complexity of the model. In an infinite-dimension parameter space, a Bayesian nonparametric model uses only a finite subset of the available dimensions to explain a sample of observations, with the complexity of the model adapting to the sample data.

C-statistic: A measure of the discriminative ability of a logistic regression model. The concordance (or c) statistic is a unit-less index denoting the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome occur compared with a randomly selected subject who did not experience the event.

Effect modification: Occurs when the magnitude of the effect of the primary treatment or exposure on an outcome differs depending on the level of a third variable (e.g., patient characteristics). In the presence of effect modification, the use of an overall effect estimate is inappropriate.

Ensemble learning: A type of machine learning approach that combines multiple learning algorithms or models to predict an outcome to obtain better model performance than any of the individual models.

Genome-wide association study (GWAS): An observational study of a genome-wide set of genetic variants in different individuals to examine associations with variants with an outcome or trait.

Heterogeneous treatment effects (HTE): Nonrandom variability in the direction or magnitude of a treatment effect, measured using clinical outcomes. HTE is fundamentally a scale-dependent concept and therefore, for clarity, the scale should generally be specified.

- **Clinically important HTE:** Occurs when variation in the risk difference across patient subgroups span an important decision threshold, which depends on treatment burden (including treatment-related harms and costs). It is generally assessed on the absolute scale.

- **Predictive HTE analysis:** The main goal of predictive HTE analysis is to develop models that can be used to predict which of two or more treatments will be better for a particular individual.

  — **Risk modeling approach:** An approach to predictive HTE analysis in which a multivariable model that predicts the risk of an outcome (usually the primary study outcome) is applied to disaggregate patients in trials to examine risk-based variation in treatment effects.

    ▫ **External models versus endogenous/internally derived models:** An external risk model has been developed from an external trial or cohort population that can be applied for HTE analysis of the trial. An endogenous or "internal" risk model is one developed directly on the trial population that does not include a term for treatment assignment.

  — **Treatment effect modeling approach:** An approach to predictive HTE analysis that develops a model directly on randomized trial data to predict treatment effects (i.e., the contrast in outcome risks under two alternative treatment conditions). Unlike risk modeling, the model incorporates a term for treatment assignment and permits the inclusion of treatment-by-covariate interaction terms.

Net benefit: A decision analytic measure that puts benefits and harms on the same scale. This is achieved by specifying an exchange rate based on the relative

value of benefits and harms associated with interventions. The exchange rate is related to the probability threshold to determine whether a patient is classified as being positive or negative for a model outcome, or (when applied to trial analysis) as being treatment-favorable versus treatment-unfavorable.

Overfitting: A key threat to the validity of a model when predictions do not generalize to new subjects outside the sample under study. Overfitting occurs when a model conforms too closely to the idiosyncrasies or "noise" of the limited data sample on which it is derived.

Penalized regression: A set of regression methods, developed to prevent overfitting, in which the coefficients assigned to covariates are penalized for model complexity. Penalized regression is sometimes referred to as shrinkage or regularization. Examples of penalized regression include LASSO, ridge, and elastic net regularization.

Predictive analytics: The field of predictive analytics encompasses a variety of statistical methods including prediction modeling, machine learning, and data mining techniques to make use of existing data to predict future events.

Reference class: A group of similar cases that is used to make predictions for an individual case of interest. The "reference class problem" refers to the fact that there are an indefinite number of different ways to define similarity.

Regression tree-based methods: Algorithms that use a recursive partitioning approach to predict categorical (classification tree) or continuous outcomes (regression tree).

Subgroup analysis: An analysis that examines whether specific patient characteristics modify the effects of treatment on an outcome.

Appendix B

WORKSHOP PARTICIPANTS, WEB PARTICIPANTS, AND STAFF

WORKSHOP PARTICIPANTS

Derek Angus, M.D., M.P.H., F.R.C.P.
Chair of Critical Care Medicine
University of Pittsburgh School of Medicine

Katrina Armstrong, M.D.
Chair
Department of Medicine
Physician in Chief
Massachusetts General Hospital

Naomi Aronson, Ph.D.
Executive Director
Clinical Evaluation, Innovation, and Policy
Blue Cross Blue Shield Association

Anirban Basu, Ph.D., M.S.
Director
Comparative Health Outcomes, Policy, and Economics Institute
University of Washington

Sanjay Basu, M.D., Ph.D.
Assistant Professor of Medicine
Stanford University

Jesse Berlin, Sc.D.
Vice President
Global Head of Epidemiology
Johnson & Johnson

Arlene Bierman, M.D., M.S.
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research & Quality

Cynthia Boyd, M.D., M.P.H.
Professor
Department of Geriatric Medicine and Gerontology
School of Medicine
Johns Hopkins University

James Burke, M.D., M.S.
Associate Professor of Neurology
University of Michigan

Ben van Calster, Ph.D.
Assistant Professor
Department of Development and Regeneration
University of Leuven

Thomas Concannon, Ph.D., M.A.
Senior Policy Researcher
RAND Corporation

Jennifer Croswell, M.D., M.P.H.
Senior Program Officer
Research Synthesis
Patient-Centered Outcomes Research Institute

William Crown, Ph.D.
Chief Scientific Officer
OptumLabs

John Cuddeback, M.D., Ph.D.
Chief Medical Informatics Officer
American Medical Group Association

Ralph D'Agostino, Sr., Ph.D.
Professor
Mathematics, Biostatistics, and Epidemiology
Boston University

Frank Davidoff, M.D.
Editor-in-Chief (Emeritus)
Annals of Internal Medicine

Karina Davidson, Ph.D., M.A.Sc.
Vice Dean
Organizational Effectiveness
Columbia University

Robert Dubois, M.D., Ph.D.
Chief Science Officer
Executive Vice President
National Pharmaceutical Council

Emily Evans, Ph.D., M.P.H.
Program Officer
Science
Patient-Centered Outcomes Research Institute

Robert Golub, M.D.
Deputy Editor
Journal of the American Medical Association

Steve Goodman, M.D., M.H.S., Ph.D.
Associate Dean for Clinical and Translational Research
School of Medicine
Stanford University

Sheldon Greenfield, M.D.
Executive Co-Director
Health Policy Research Institute
School of Medicine
University of California, Irvine

David Grossman, M.D., M.P.H.
Senior Investigator
Kaiser Permanente Washington Health Research Institute

John Haaga, Ph.D.
Director
Division of Behavioral and Social Research
National Institute on Aging

Frank Harrell, Ph.D.
Professor of Biostatistics
School of Medicine
Vanderbilt University

Rodney Hayward, M.D.
Director
National Clinician Scholars Program
Institute for Healthcare Policy & Innovation
Professor
Department of Internal Medicine and Department of Health Management and Policy
University of Michigan

Patrick Heagerty, Ph.D.
Chair
Department of Biostatistics
University of Washington

Erik Hess, M.D., M.Sc.
Vice Chair for Research
Department of Emergency Medicine
The University of Alabama at Birmingham

David Hickam, M.D., M.P.H.
Program Director
Patient-Centered Outcomes Research Institute

Ralph Horwitz, M.D., M.A.C.P.
Professor Emeritus of Medicine and Epidemiology
School of Medicine
Yale University

Stanley Ip, M.D.
Associate Director
Science
Patient-Centered Outcomes Research Institute

A. Cecile J. W. Janssens, Ph.D., M.Sc., M.A.
Professor of Epidemiology
Rollins School of Public Health
Emory University

Sherrie Kaplan, Ph.D., M.P.H.
Assistant Vice Chancellor for Healthcare Measurement and Evaluation
University of California, Irvine

David Kent, M.D., M.S.
Director
Predictive Analytics and Comparative Effectiveness Center
Tufts Medical Center

David van Klaveren, Ph.D., M.Sc.
Assistant Professor
Medical Statistics
Leiden University Medical Center
Tufts Medical Center

Michael Kurilla, M.D., Ph.D.
Director of Clinical Innovation
National Center for Advancing Translational Sciences

Fan Li, Ph.D.
Associate Professor of Statistical Science
Duke University

Hester Lingsma, Ph.D.
Associate Professor Medical Decision Making
Erasmus University Medical Center, Rotterdam

Jeanne Mandelblatt, M.D., Ph.D.
Professor of Medicine and Oncology
Georgetown University

Anne-Marie Mazza, Ph.D., M.A.
Director of Strategic Initiatives
National Academy of Medicine

Newell McElwee, Pharm.D., M.S.P.H.
Vice President
Health Economics and Outcomes Research
Boehringer Ingelheim Pharmaceuticals

Seth Morgan, M.D.
District Activist Leader
National Multiple Sclerosis Society

Sally Morton, Ph.D.
Dean of the College of Science
Virginia Tech

Jason Nelson, M.P.H.
Statistician
Predictive Analytics and Comparative Effectiveness Center
Tufts Medical Center

Peter Neumann, Sc.D.
Director
Center for the Evaluation of Value and Risk in Health
Tufts Medical Center

Bray Patrick Lake, M.F.S.
Director of Stakeholder Engagement and the Research Together Program
Duke Clinical Research Institute

Michael Pencina, Ph.D.
Vice Dean for Data Science and Information Technology
Duke Clinical Research Institute

Josh Peterson, M.D., M.P.H.
Associate Professor of Biomedical Informatics and Medicine
Vanderbilt University

Joseph Ross, M.D., M.H.S.
Associate Professor of Medicine and Public Health
School of Medicine
Yale University

Darshak Sanghavi, M.D.
Chief Medical Officer
Senior Vice President of Translation
OptumLabs

Joseph Selby, M.D., M.P.H.
Executive Director
Patient-Centered Outcomes Research Institute

Harry Selker, M.D., M.S.P.H.
Dean of the Tufts Clinical and Translational Science Institute
Tufts University

Nilay Shah, Ph.D.
Director for Research
Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery
Mayo Clinic

Changyu Shen, Ph.D.
Associate Professor
Smith Center for Outcomes Research in Cardiology
Beth Israel Deaconess Medical Center

Hal Sox, M.D.
Program Director
Science
Patient-Centered Outcomes Research Institute

John Spertus, M.D., M.P.H., F.A.C.C.
Chair
Professor of Medicine
University of Missouri–Kansas City

Christine Stake, D.H.A.
Research Operations Manager
Ann & Robert H. Lurie Children's Hospital of Chicago

Ewout Steyerberg, Ph.D.
Professor
Clinical Biostatistics and Medical Decision Making
Leiden University Medical Center

Jeremy Sussman, M.D., M.S.
Assistant Professor
Department of Internal Medicine
University of Michigan Medical School

Robert Temple, M.D.
Deputy Center Director for Clinical Science Center for Drug Evaluation and Research
U.S. Food and Drug Administration

Ravi Varadhan, Ph.D.
Associate Professor of Oncology
Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

Benjamin Wessler, M.D., M.S.
Director
Valve Center
Assistant Professor
Predictive Analytics and Comparative Effectiveness Center
Tufts Medical Center

Evelyn Whitlock, M.D., M.P.H.
Chief Science Officer
Patient-Centered Outcomes Research Institute

Richard Willke, Ph.D.
Chief Science Officer
International Society for Pharmacoeconomics and Outcomes Research

John Wong, M.D.
Chief
Division of Clinical Decision Making
Tufts Medical Center

WEB PARTICIPANTS

John Ioannidis, M.D., D.Sc.
C.F. Rehnborg Chair in Disease Prevention
Stanford University

Michael Rothberg, M.D., M.P.H.
Vice Chair for Research
Medicine Institute
Cleveland Clinic

Richard Kravitz, M.D., M.S.P.H.
Co–Vice Chair for Research
Department of Internal Medicine
University of California, Davis

Navdeep Tangri, M.D., Ph.D.
Associate Professor
Division of Nephrology
University of Manitob

STAFF

Noor Ahmed, M.Eng.
Research Associate
National Academy of Medicine

Robert Pool, Ph.D.
Science Writer
Hired Pens LLC

Jennifer Lutz, M.A.
Program Coordinator II
Predictive Analytics and Comparative Effectiveness Center
Tufts Medical Center

Danielle Whicher, Ph.D., M.H.S.
Senior Program Officer
National Academy of Medicine

Jessica Paulus, Sc.D.
Assistant Professor
Predictive Analytics and Comparative Effectiveness Center
Tufts Medical Center

*All affiliations and titles were accurate at the time of the workshop.

Appendix C

WORKSHOP AGENDA

Evidence and the Individual Patient: Understanding Heterogeneous Treatment Effects for Patient-Centered Care

May 31, 2018
National Academy of Sciences Building
Lecture Room
2101 Constitution Avenue, NW
Washington, DC 20001

Meeting Focus: *Leveraging data to examine heterogeneous treatment effects to personalize and improve patient care*

Motivating Questions:

1. *Potential:* How can clinical trial data be analyzed to yield reliable patient-centered treatment effect estimates? What are the state-of-the-science methods for assessing treatment heterogeneity?
2. *Risks:* How can we be sure personalizing evidence will improve decision making, as compared with the default of relying on overall average treatment effects? What are the evidentiary standards for implementing changes to clinical practice to personalize care based on evidence of heterogeneous treatment effects?
3. *Lessons learned:* What can be learned from the challenges of genomics-based personalized medicine? What can be learned from the efforts of previous clinical trialists to understand more personalized treatment effect estimates?
4. *Strategies:* How should clinical research and clinical practice be redesigned to support the generation and the dissemination of patient-centered evidence?

continued

Outcomes Anticipated: The conference will stimulate discussion and further collaborative action to advance the research and policy agenda for patient-centered evidence and will inform the development of a white paper outlining the optimal methodological approaches to personalizing treatment effects, and the clinical contexts in which these approaches are likely to be of most value.

8:30 a.m. Coffee and light breakfast available

9:00 a.m. Welcome, Introductions, and Workshop Overview

Welcome from the National Academy of Medicine
Anne-Marie Mazza, National Academy of Medicine

Opening Remarks and Workshop Overview
Joe Selby, Patient-Centered Outcomes Research Institute (PCORI)

9:15 a.m. Overview of Heterogeneous Treatment Effects: Moving from Evidence-Based Medicine to Personalized/ Precision Medicine

Speakers will present a conceptual overview of heterogeneous treatment effects, as well as examples of clinical trials analyzed to yield more personalized treatment effect estimates. Discussion will focus on how changes in the design of clinical research might enable a better understanding of how treatment effects vary across individuals.

Moderator: *Harry Selker,* Tufts Medical Center

Speakers:
David Kent, Tufts Medical Center
Sanjay Basu, Stanford University
Derek Angus, University of Pittsburgh

Discussants:
Sheldon Greenfield, University of California, Irvine
Bob Temple, U.S. Food and Drug Administration

Q&A and Open Discussion

10:45 a.m. Break

11:00 a.m. An Equation-Free Presentation of New Methods for Prediction of Treatment Benefit and Model Evaluation

This session will focus on statistical methods. Speakers will discuss lessons learned from the genomics revolution, machine learning methods for the analysis of trial data, and new methods for evaluating models that predict treatment benefit.

Moderator: *Ewout Steyerberg,* Leiden University Medical Center

Speakers:

A. Cecile J.W. Janssens, Emory University Rollins School of Public Health
Fan Li, Duke University
Patrick Heagerty, University of Washington School of Public Health

Discussants:

Frank Harrell, Vanderbilt University
Michael Pencina, Duke Clinical Research Institute

Q&A and Open Discussion

12:20 p.m. Break

Participants will pick up lunch.

12:35 p.m. Discussion with Stakeholders

This session will focus on how representatives of several patient communities have applied research to guide their own care, given their own individual circumstances. Additional stakeholders will contribute to the discussion of how to better align evidence with patient-centered care.

Moderator: *Bray Patrick-Lake,* Duke University

Panelists:

Thomas Concannon, RAND Corporation

Seth Morgan, National Multiple Sclerosis Society, Advocate and Patient Stakeholder
Christine Stake, Ann & Robert H. Lurie Children's Hospital of Chicago, Patient Stakeholder

Reactors:
Robert Dubois, National Pharmaceutical Council
Karina Davidson, Columbia University

Q&A and Open Discussion—Engagement with other stakeholders

1:30 p.m. Break

1:45 p.m. From Research into Practice: Implementation and Oversight

This session will focus on barriers to implementation applying predictions in clinical practice and how to overcome these barriers. Speakers will discuss how to go beyond "all-or-nothing" quality measures to incentivize more personalized care.

Moderator: *Nilay Shah,* Mayo Clinic

Speakers:
John Spertus, Saint Luke's Mid America Heart Institute
Josh Peterson, Vanderbilt University
Rodney Hayward, University of Michigan

Discussants:
Naomi Aronson, Blue Cross Blue Shield Association
Katrina Armstrong, Massachusetts General Hospital

Q&A and Open Discussion

3:15 p.m. Opportunities for Collaborative Action

The aim of this session is to reflect on key themes from the day's discussion, focusing on innovative methods for understanding heterogeneous treatment effects, challenges related to implementation and oversight to personalize care,

and outstanding policy and research questions that need to be addressed to accelerate progress.

Moderator: *Joe Selby,* PCORI

Reactors:
Steven Goodman, Stanford University
Evelyn Whitlock, PCORI

Closing Remarks:
David Kent, Tufts Medical Center

4:30 p.m. Adjourn

Workshop Planning Committee

David M. Kent (*Chair*), M.D., M.S., Tufts Medical Center
Thomas Concannon, Ph.D., M.A., RAND Corporation
Robert Golub, M.D., *Journal of the American Medical Association*
Sheldon Greenfield, M.D., University of California, Irvine, School of Medicine
Rodney Hayward, M.D., University of Michigan
A. Cecile Janssens, Ph.D., M.Sc., M.A., Emory University Rollins School of Public Health
Muin J. Khoury, M.D., Ph.D., Centers for Disease Control and Prevention
Peter Rothwell, M.D., Ph.D., University of Oxford
Ewout Steyerberg, Ph.D., Leiden University Medical Center
Andrew J. Vickers, D.Phil., Memorial Sloan Kettering Cancer Center