Lecture Notes in Statistics 1205
Proceedings

Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials

Springer

Lecture Notes in Statistics-Proceedings 1205

Edited by
P. Bickel
P.J. Diggle
S.E. Fienberg
U. Gather
I. Olkin
S. Zeger

For further volumes:
http://www.springer.com/series/8440

Thomas R. Fleming • Bruce S. Weir
Editors

Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials

Springer

Editors
Thomas R. Fleming
Department of Biostatistics
University of Washington
Seattle, WA 98195-7232, USA

Bruce S. Weir
Department of Biostatistics
University of Washington
Seattle, WA 98195-7232, USA

ISSN 1869-7240
ISBN 978-1-4614-5244-7 ISBN 978-1-4614-5245-4 (eBook)
DOI 10.1007/978-1-4614-5245-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012952276

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Continuing the tradition, once every half decade, of hosting a major biostatistical meeting, where thought leaders can gather to address scientific issues of current and compelling importance, the organizing committee and sponsors held the Fourth Seattle Symposium on Biostatistics on November 22 and 23, 2010. The topic area for this successful meeting was clinical trials, with focus on the use of biomarkers, issues in multi-regional clinical trials, and identifying and addressing safety signals. The event was sponsored by Axio Research, Bristol-Myers Squibb, Genentech, Novartis and Onyx and was co-sponsored by the UW School of Public Health and the Division of Public Health Sciences at the Fred Hutchinson Cancer Research Center (FHCRC). The symposium featured keynote lectures by Robert O'Neill, Ross Prentice and Robert Temple, as well as invited talks by Jesse Berlin, Christy Chuang-Stein, David DeMets, Bill DuMouchel, Susan Ellenberg, Thomas Fleming, Laurence Freedman, Margaret Pepe, Steve Self, Richard Simon, Bruce Weir, John Whittaker and Janet Wittes. Invited panelists included Jesse Berlin, Bruce Binkowitz, Christy Chuang-Stein, Bill DuMouchel, Susan Ellenberg, Thomas Fleming, Henry Fuchs, Dominic Labriola, Robert O'Neill, Robert Temple and Janet Wittes. There were 200 attendees at the symposium. In addition, more than 100 people attended short courses delivered on November 20 and 21, 2010. At these short courses, "Statistical Design of Sequential Clinical Trials in R" was taught by Scott Emerson and Dan Gillen, "The Use of Genetic Marker Data in Clinical Trials" was taught by Bruce Weir and Patrick Heagerty, "Data Monitoring Committees: A Practical Approach" was taught by Susan Ellenberg, Thomas Fleming and David DeMets, "Statistical Evaluation of Markers for Classification and Prediction" was taught by Margaret Pepe and "Practice Issues in the Conduct and Reporting of Large-Scale Clinical Trials: The Women's Health Initiative Experience" was taught by Garnet Anderson and Andrea LeCroix.

When the UW School of Public Health was formed in 1970, biostatistics as a discipline was very young. In the subsequent 40 years, both the field and the UW Department of Biostatistics have evolved in many exciting ways. The department had only seven faculty when it moved from the School of Medicine to the new School of Public Health and Community Medicine in 1970. The faculty roster

currently lists 49 regular and research faculty and 34 adjunct and affiliate faculty. Ed Perrin was the Department Chair in 1970, succeeded by Donovan Thompson, Norman Breslow, Thomas Fleming and presently Bruce Weir. The faculty have been actively involved in methodological and collaborative research in addition to graduate teaching. The choice of *Clinical Trials* as the theme for the *Fourth Seattle Symposium in Biostatistics* was a tribute to the significant contributions made by the UW and FHCRC faculty to this important area of statistical science.

The Symposium Organizing Committee consisted of Susan Ellenberg, Scott Emerson, Nathalie Ezzet, Thomas Fleming (Chair), Henry Fuchs, Lee Hooks, Dominic Labriola, Michael Ostland, Ross Prentice and Bruce Weir. The staff of the Department of Biostatistics, especially Sandra Coke**,** provided great administrative support to the symposium. The UW School of Public Health Dean Howard Frumkin, the Department Chair Bruce Weir and the Organizing Committee Chair Thomas Fleming delivered the opening remarks. The scientific sessions were chaired by Bruce Weir, Scott Emerson, Thomas Fleming, Lee Hooks, Henry Fuchs, Michael Ostland, Susan Ellenberg, Nathalie Ezzet and Dominic Labriola. We are grateful to the aforementioned people as well as all the speakers and participants for making the symposium a great success.

This volume contains most of the papers presented at the symposium, as well as some of the science presented at the short courses. These papers encompass recent methodological advances on several important topics, summaries of the state of the art of methodology in key areas of clinical trials, as well as innovative applications of the existing theory and methods. This collection serves as a reference for those working in several key areas of clinical trials.

Each of the 12 papers in this volume was referred by two or three peer reviewers, and their comments were incorporated by the authors into the final versions of the papers. The referees are listed at the end of this book. We are indebted to them for their time and efforts. We also appreciate the guidance and assistance by Marc Springer of Springer-Verlag.

Contents

Part I
Biomarkers: Role in the Design and Interpretation of Clinical Trials

The Role and Potential of Surrogate Outcomes in Clinical Trials: Have We Made Any Progress in the Past Decade?

David L. DeMets

Abstract Randomized clinical trials are the standard method for evaluating new interventions or comparing existing ones. Trials which use clinical outcomes as the primary outcome can be large, require lengthy follow-up, and can be expensive. For these reasons, researchers have sought to use intermediate outcomes such as biomarkers as a substitute or surrogate for the clinical outcome. Over a decade ago, this practice had become common. Fleming and DeMets (Ann Intern Med 125:605–613, 1996) reported many cases where the use of a biomarker as a surrogate outcome failed to reliably assess the effect of the intervention, in some cases missing harmful effects including mortality. Recently, the Institute of Medicine (IOM) reviewed the state of the art and came to similar conclusions that biomarkers have often proved to be unreliable as a surrogate [Committee on Qualifications of Biomarkers and Surrogate Endpoints in Chronic Disease, Michael C, Ball J (eds) (2010) Evaluation of biomarkers and surrogate endpoints in chronic disease. National Academies Press, Washington]. They proposed that biomarkers must meet certain criteria including analytic validity, strong correlation with the clinical outcome and the ability to capture the full effects of the intervention. The use of a biomarker as a surrogate must be done so in the context of its intended use, and done so with great caution. While the IOM report further clarifies the necessary requirements of a potential biomarker as a surrogate, the report still recommends caution in using surrogate outcomes in final phases of intervention evaluation as did Fleming and DeMets (Ann Intern Med 125:605–613, 2004).

D.L. DeMets (✉)
Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, 610 Walnut Street, Madison, WI, 53726 USA
e-mail: demets@biostat.wisc.edu

T.R. Fleming and B.S. Weir (eds.), *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, Lecture Notes in Statistics 1205,
DOI 10.1007/978-1-4614-5245-4_1,

1 Introduction

Clinical trials have been the primary method for evaluating new interventions or strategies for disease diagnosis, prevention, and treatment over the past four decades. These interventions may be drugs, biologics, devices, procedures, and dietary or behavioral modifications. While the ultimate test of a new intervention would be the modification of a clinically important outcome, trials with such a design can be large, lengthy, and costly. Attempts to improve trial efficiency have included a substitute outcome for the clinically important outcome that may be easier, cheaper, or quicker to measure and may also result in a smaller trial. The intermediate outcomes or biomarkers used for this purpose are often referred to as surrogate outcomes, defined as a biomarker that is intended to be used as a substitute for a clinical outcome in evaluating a new intervention [1].

Examples of such intermediate markers that have been used previously in the evaluation of new interventions are blood pressure and cholesterol levels as a substitute for cardiovascular events such as death or nonfatal myocardial infarction. While the use of surrogate outcomes has become common in recent years, in 1996 Fleming and DeMets reviewed the experience at that point as to whether or not the use of biomarkers as surrogates in clinical trials had proven to be reliable [1]. Their conclusion was that there were many examples in several disciplines where the use of biomarkers as surrogate outcomes had been misleading with regard to benefit and risk for new interventions. Recently, the Institute of Medicine (IOM) reviewed the same subject in the context of use of biomarkers as surrogates for nutritional intervention claims and came to similar conclusions [2]. In addition, the IOM report presents a structure for the evaluation of a biomarker for potential use as a surrogate outcome. This paper will provide a brief overview of the requirements for a biomarker to be a valid surrogate with some early and more recent examples of biomarkers not being reliable as surrogates.

2 Basic Requirements for a Potential Surrogate

In 2010, the Institute of Medicine (IOM) of the National Academies issued a report "Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease" which reviewed the requirements for a potential biomarker to be used as a surrogate end point [2]. While the initial stimulus for this evaluation was for nutritional biomarkers as a surrogate for health claims, the IOM report provided a general structure for evaluation of any biomarker for such use. Biomarkers measure some biological process and include, for example, not only physiological and blood measurements but also genetic or genomic signatures. Biomarkers can be used to describe risk, risk exposure, or intermediate response to an intervention or as surrogates or substitutes for clinical outcomes in evaluating a new intervention. The report argues that the Food and Drug Administration (FDA) and other regulatory agencies should use the same degree of scientific rigor across all categories of products including drugs,

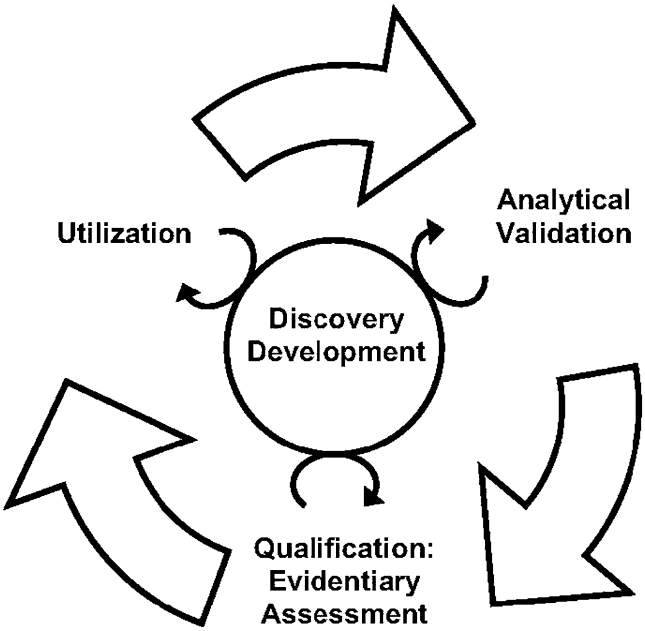

Fig. 1 The steps of the evaluation framework are interdependent. While a validated test is required before qualification and utilization can be completed, biomarker uses inform test development, and the evidence suggests possible biomarker uses. In addition, the circle in the center signifies ongoing processes that should continually inform each step in the biomarker evaluation process

biologics, devices, foods, and supplements. The evaluation process for a biomarker to be a surrogate should have three steps (1) analytical validation, (2) qualification, and (3) utilization (Fig. 1).

For analytic validation, the IOM report defined this as an assessment of the measurement assay and performance [2]. They recommended that a biomarker must be evaluated for its limit of detection, limit of quantification, reliability, and reproducibility across different laboratories. In addition, to be a surrogate outcome, a biomarker must have adequate sensitivity and specificity. Sources of variability in the biomarker measurement must be understood and controlled to the extent possible through good clinical laboratory and quality control practices. While similar to routine laboratory validation, according to the IOM report, there is still no uniform set of criterion for biomarker assay validation [2].

The second step of qualification of a biomarker as a surrogate has two parts (1) demonstrating a correlation between the biomarker and the true clinical outcome and (2) demonstrating that the total clinical impact of the intervention is captured by the biomarker. Prentice [3] described in detail these statistical criteria. The first requirement is easy to understand but a common misconception is that if a biomarker is highly correlated with a true clinical outcome, it can be used as a surrogate. However, as stated by Fleming and DeMets, "a correlate does not make a surrogate" [1]. The second criterion is a necessary and much stronger condition

than just correlation. Figure 2 illustrates the simple conception of a biomarker as a surrogate. The disease process can be measured by a biomarker which is in the causal pathway to the true clinical outcome. Correlation between the biomarker and the clinical outcome would be very high in this scenario. If this were the case, then any intervention on the disease process which modified the biomarker would capture the clinical effect. However, biology is almost surely more complicated than this simple pathway. Figure 3 depicts three alternative and more complicated pathways. These are only a sample of potential pathways. In Fig. 3, the biomarker would be correlated to some degree with the clinical outcome in all cases. However, there is not a single or direct pathway between the biomarker and the clinical outcome. For example, in case A, the biomarker is not at all in the causal pathway but affected simultaneously with the clinical outcome by the disease process, thus producing a high correlation that has no associated causation. Any intervention could have a profound effect on the biomarker but have absolutely no impact on the clinical outcome. In case B, the biomarker is in the causal pathway but there is at least one other pathway where the biomarker plays no role. An intervention might affect the biomarker and have some degree of impact on the clinical outcome, but it could be overwhelmed by the other pathway. For example, the intervention might have the desired effect on the biomarker but have a negative or harmful effect through the other pathway that would not be identified if only the biomarker was measured. In case C, an intervention might have no effect on the biomarker but still have a direct effect through another pathway on the clinical outcome. In this case, a partially effective intervention might be totally missed. A great deal of statistical research has been focused on this issue of what criteria is adequate for a biomarker to be a clinical surrogate but the criteria set forth by Prentice captures the essence of the challenge [4–17]. That is, the biomarker must be highly correlated with the clinically relevant outcome and must capture all, or nearly all, of the effect of the intervention on the clinical outcome. While these criteria are challenging to meet, failure to meet those requirements can lead to interventions becoming part of clinical practice and yet have no or even harmful effects as illustrated by examples described below.

Even if Fig. 2 were the correct pathway, the effect of the intervention seen in the biomarker could be misleading. The effect could be underestimated if there is considerable noise or measurement error in the biomarker. Alternatively, the effect could be overestimated if the effect on the biomarker is of sufficient size to produce a meaningful clinical effect.

The third step in the IOM recommendation is to take into consideration the intended use of the biomarker as a surrogate. For example, a biomarker might be suitable for identifying disease risk without any reference to intervention effect. That is, the correlation of the biomarker with the clinical outcome might be sufficiently high to have utility in describing risk. A biomarker might also be a useful intermediate or surrogate in the drug or device development process. For example, in developing a new drug, device, or intervention to reduce cardiovascular risk, assessing the impact on reducing blood pressure might be useful to rule out interventions which do not have this intended effect. However, as indicated above, other non-intended affects might not be captured by total reliance on this biomarker.

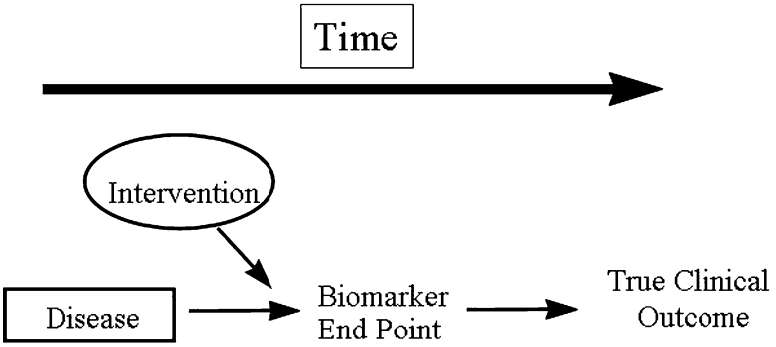

Fig. 2 Common misconception of a causal pathway and a biomarker. The setting that provides the greatest potential for the surrogate endpoint to be valid. Reprinted from Ann Intern Med 1996; 125:605–613

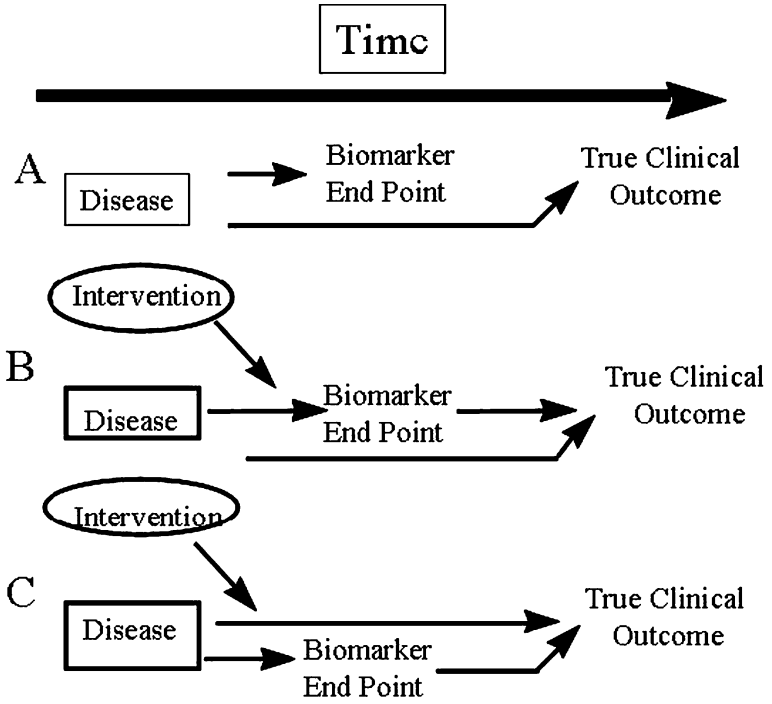

Fig. 3 Possible causal pathways relative to a biomarker. Reasons for failure of biomarker end points: (*A*) The biomarker is not in the causal pathway of the disease process; (*B*) Of several causal pathways of disease, the intervention affects only the pathway mediated through the biomarker; (*C*) The biomarker is not the pathway of the intervention's effect or is insensitive to its effect

3 Some Earlier Examples

While there are examples of a biomarker being successful in risk assessment or intervention development, there are many examples of failures when a biomarker was relied on as a valid surrogate. Unfortunately, we may not always have the complete clinical outcome data to assess whether the biomarker was indeed a valid surrogate. In addition, an example where the biomarker captures the effect for one specific intervention does not guarantee that it will be reliable for the next intervention of a different class or type, or even a variation within the same class. We shall briefly summarize a few of the examples described by Fleming and DeMets [1] and by the IOM report [2]. What is so remarkable is that these cases of biomarker failure can be found across a wide variety of disease areas, across a wide variety of interventions and even within a class of interventions where one biomarker success did not translate into other interventions even of the same class.

Perhaps one of the most dramatic and important examples of a biomarker failure as a surrogate outcome comes from the field of cardiology and the use of arrhythmia suppressing drugs to reduce death from cardiovascular complications. Observational data had shown that ventricular arrhythmias were shown to have as much as a four-fold increase in cardiovascular death [18, 19]. This observation led to the arrhythmia suppression hypothesis that reducing cardiac arrhythmias would reduce sudden death. Drugs were developed to suppress these arrhythmias and approved by the FDA for use in high-risk patients. The assumption was that, in general, suppressing these arrhythmias would result in a reduction of cardiovascular death and in fact some drugs with this arrhythmia suppressing effect began to be used beyond just the highest risk patients. That is, many clinicians practiced as if biomarker arrhythmia suppression would be an adequate surrogate for survival. The Cardiac Arrhythmia Suppression Trial (CAST) was a double-blind placebo-controlled trial of three such drugs (encainide, flecainide, and moricizine) to test the hypothesis that these drugs would reduce the risk of cardiovascular death in patients with a recent myocardial infarction and at least 10 premature ventricular beats per minute [18]. Each drug had a matching placebo. In order to be eligible, a patient had to have a suppressible arrhythmia as determined in a run-in phase. Initially, the data monitoring committee was blinded to drug assignment. The moricizine arm got a late start so the data for the other two arms were monitored initially by a data monitoring committee blinded to treatment assignment. Early trends in mortality were assumed to be beneficial, as expected, based on clinical practice and a belief that arrhythmia suppression was a valid surrogate. When these early trends became stronger, the data monitoring committee was unblinded and startled to learn that the trends were contrary to expectation. At the time the data monitoring committee recommended trial termination, as 56 deaths were observed on the two drug arms compared to 22 in the matching placebo arm. When the follow-up data for these randomized patients was completed, there were 63 deaths on the two drug arms compared to 26 in the placebo arm [19]. Similarly, there were 43 sudden deaths on drugs and 16 on placebo. Later, when the results for the moricizine arm became available, this arm was terminated as well

with an increased risk [20]. In this third arm, the brief run-in period was modified after the results of the other two drugs arms became available to be randomized to either moricizine or placebo. If patient's arrhythmias were suppressed during the short run-in period, they were randomized to drug or placebo as before. Even exposure to moricizine during the run-in period demonstrated a strong trend for increased risk compared to the placebo. This case has several important lessons. First, the observational data was convincing about arrhythmia suppression as a surrogate and the biology seemed plausible. Assumptions were made about the clinical effect. A large number of patients were exposed to these very harmful drugs. Furthermore, the second half of CAST demonstrated that even a short exposure to moricizine was risky. This suggests that clinicians in their normal practice could not detect this increased risk in a group of patients who were already at risk due to a prior heart attack. Before these drugs were used routinely in clinical practice, the definitive test using clinical outcomes, in retrospect, should have been done.

While the lessons from CAST are dramatic, this is not a unique example in cardiology for this particular patient diagnosis. Other drugs such as quinidine and lidocaine with known arrhythmia controlling activity were shown to have increased mortality risk [21–23].

4 Lipid-Lowering Interventions

The Framingham Heart Study identified that high cholesterol levels including low density lipids (LDL) were associated with increased cardiovascular mortality [24]. Strategies to lower lipid values were identified including drug interventions such as niacin and clofibrate. The Coronary Drug Project (CDP) was a multi-armed randomized placebo-controlled trial of these two drugs as well as high and low doses of estrogen, also known to reduce cholesterol levels, to test the lipid lowering hypothesis [25]. The CDP was a trial with a planned 7 years of follow-up with death and death from coronary heart disease as the primary outcomes. The two estrogen arms were terminated early with increased cardiovascular risk probably due to increased clotting risk. Neither the niacin or clofibrate arm reduced total mortality although there was a favorable trend for niacin [25]. Several other lipid-lowering trials combined in a meta-analysis did not show a reduction in total mortality but actually had an increase in noncardiovascular death which offset a reduction in cardiovascular death [26]. Despite the consistent correlation between high cholesterol levels and increased risk, this reduction of serum lipid values was not an adequate biomarker to be used as a valid surrogate.

While several interventions existed which reduce cholesterol, the clinical benefit was not demonstrated until the Scandinavian Simvastatin Survival Study (4S) was done [27]. The multicenter randomized double-blind placebo-controlled trial evaluated one of the statins and observed a 25% reduction in cholesterol with a 30% reduction in total mortality. This trial has had a major impact on clinical practice. However, just because one statin had a beneficial effect does not mean

that cholesterol lowering with a statin qualifies that as a surrogate. One statin trial (Baycol) was terminated early with an increase in mortality [28]. In another study, the trial ILLUMINATE evaluated a member of a new class of drugs, torcetrapib, that decreased LDL cholesterol (the bad cholesterol) and increased HDL cholesterol (the good cholesterol) but was terminated early because of an increase in death and cardiac events [29].

Several large epidemiological studies demonstrated a strong correlation between hormone replacement treatment (HRT) usage, either estrogen or estrogen-progestin supplementation, and a decreased risk in cardiovascular risk [30, 31]. HRT is known to reduce cholesterol levels and reduce bone density loss in post-menopausal women. Estrogen alone is used for women with a hysterectomy and an estrogen–progestin combination for women with an intact uterus. HRT supplementation became one of the most widely prescribed medicines, assuming that cholesterol reduction was a surrogate for cardiovascular risk. The Women's Health Initiative (WHI) was a factorial trial evaluating the impact of low-fat diets, hormone replacement therapy (HRT), and calcium supplementation [32, 33]. The HRT component was actually two trials, one comparing estrogen and placebo in women without a uterus and the other comparing estrogen–progestin with placebo in women with an intact uterus, with clinical outcomes of cardiovascular death or cardiovascular events such as nonfatal heart attack or nonfatal stroke. As expected, HRT lowered LDL cholesterol and reduced bone density loss with an accompanying reduction in major fractures. However, both trials were terminated early with increased cardiovascular risk, with clotting problems being the major issue. For the estrogen–progestin arms, there was also an increase in cases of uterine cancer. These results clearly refuted the assumption that cholesterol reduction with HRT was an adequate surrogate for cardiovascular risk. Unfortunately, millions of women were treated with HRT under the false assumption of cardiovascular benefit. Interestingly, three decades earlier, men given low and high doses of estrogen also experienced increased risk due to clotting problems even though LDL was lowered [34]. This early CDP lesson on the failure of cholesterol reduction as a surrogate was missed in later research such as in design of the WHI. While the intervention effect of lowering LDL was observed, the clotting problem was not anticipated for women taking HRT (Fig. 4).

As described by Fleming and DeMets [1], there are many other examples in cardiology involving biomarkers which failed to be reliable surrogates for cardiovascular risk. These include blood pressure lowering [35–39] for cardiovascular morbidity and exercise tolerance for congestive heart failure [40–42]. Cardiology has several classic examples of biomarker failure to be a surrogate but other disciplines have such examples as well.

Cancer is the second leading cause of mortality with an associate morbidity. Cancer treatment trials have often used the biomarker tumor shrinkage as a surrogate for clinical response in drug development, for example, in breast cancer, colon cancer, and lung cancer [43]. Responses are often categorized as a complete response (no remaining tumor visible), partial response (a 50% reduction in tumor volume), no change or progression. Tumor volume has an initial challenge of

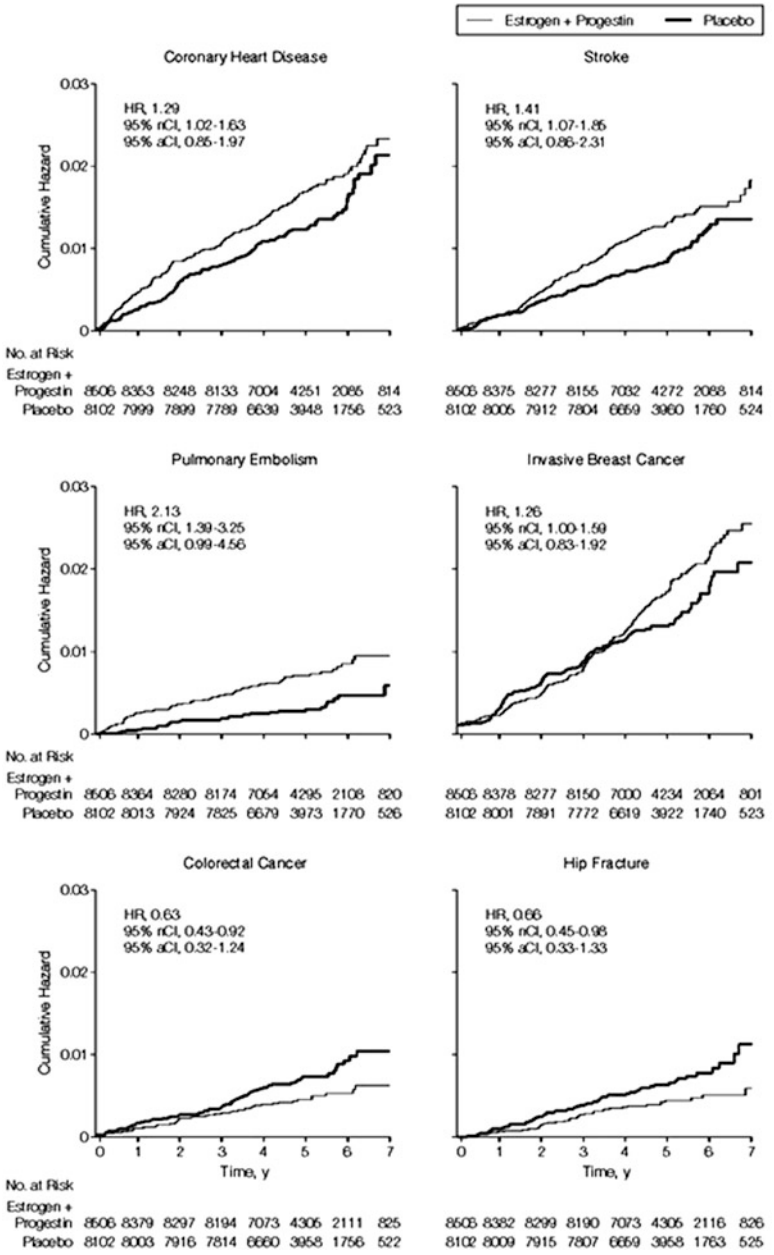

Fig. 4 (**a**) Outcomes in the Women's Health Initiative (*WHI*) trial (32) WHI Kaplan–Meier estimates of cumulative hazards for selected clinical outcomes

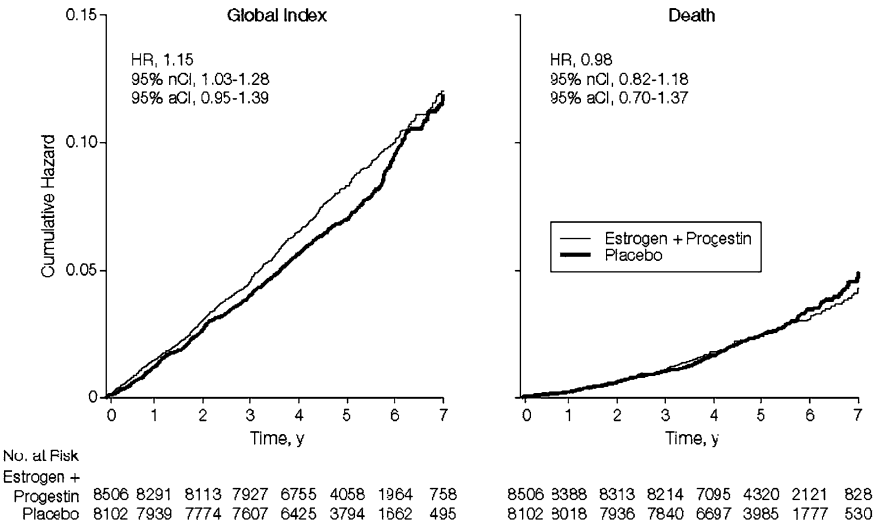

Fig. 4 (**b**)WHI Kaplan–Meier estimates of cumulative hazards for Global Index and Death. *HR* = hazard ratio; *nCI* = nominal confidence interval; *aCI* = adjusted confidence interval. JAMA 2002; 288(3):321–333

validation since measurement of tumor volume is highly variable depending on methodology. Tumor response has been widely used for drug development and early phase trials in cancer [44]. In addition, tumor volume reduction has not always been a reliable biomarker for survival [6, 45, 46]. Tumor response was also used as a surrogate for approval of cancer therapy drugs in the 1970s but later the FDA requested that a clinical survival benefit or quality of life benefit be demonstrated as well. However, with the urgency of new and effective cancer treatments, in 1996 the FDA utilized the accelerated approval process using surrogate outcomes such as progression-free survival, meaning survival with no tumor recurrence or progression [46, 47]. The assumption was that post-approval trials demonstrating a survival benefit would follow but this has not happened consistently.

A specific cancer example is the use of 5-fluorouracil plus leucovorin compared to 5-fluorouracil alone in colon cancer. The combination showed a complete plus partial response rate of 23% compared to 5-fluorouracil alone of 11%. Yet, this difference in tumor response biomarker provided no improvement in survival. These results came from a meta-analysis of over 1400 patients [48]. These disappointing results could be due to the poor effect of the combination or that the combination has other unintended adverse effects.

Other important diseases such as AIDS, osteoporosis, and infectious diseases also provide excellent examples of the failure of biomarkers [1]. Given that Fleming and DeMets discussed these problems in 1996, it is natural to ask if any progress in successful use of biomarkers as valid surrogates has been observed since. The IOM report [2] is a more recent review of this question.

5 Other Recent Examples

As discussed above, cancer is the second leading cause of mortality. Nutritional researchers observed that individuals with low levels of beta-carotene intake had a higher risk for lung cancer [49]. Beta-carotene is a plant carotenoid which is partially converted into vitamin A which is an important nutrient for many functions, including vision, gene expression, growth, and immune function. These observations were based on measures of dietary intake but were confirmed by blood level measurements. On the basis of the epidemiologic observations, the notion of increasing intake levels of beta-carotene emerged. Qualification of the biomarker, serum beta-carotene, as a surrogate outcome is itself a challenge and not given much attention[2]. Three large prevention trials were launched to test the hypothesis that increasing levels of serum beta-carotene would result in reduced risk of cancer, lung cancer in particular. [50–52].

The Alpha Tocopherol Beta-Carotene (ATBC) trial began in 1985, randomized over 29,000 Finnish smokers in a factorial trial, alpha tocopherol vs. placebo and bet-carotene vs. placebo [50]. Subjects were followed for 5–8 years using the Finnish cancer registry. Results for the beta-carotene vs. placebo component demonstrated a statistically significant increase in lung cancer incidence and lung cancer mortality for those patients receiving a beta-carotene containing study drug. While this result was unexpected, it was essentially replicated by a trial conducted in the USA, referred to as CARET [51]. CARET was a randomized placebo controlled trial of over 18,000 smokers or workers exposed to asbestos. Lung cancer mortality and coronary heart disease mortality were significantly higher for those participants on beta-carotene compared to the control arm. The relative risk for developing lung cancer was 1.28 (1.04–1.57) and 1.46 (1.07–2.0) for lung cancer death. Given that these results paralleled the ATBC results, CARET was terminated early. A third trial, the Physicians Health Study-I (PHS-I) was a randomized double-blind factorial trial of beta-carotene and aspirin compared to a matching placebo [52]. Participants were US male physicians who were largely nonsmokers. Contrary to the two other trials of smokers or workers exposed to asbestos, the PHS-I trial showed neither a benefit nor a harm with a relative risk of 0.98. In addition, there were no trends for benefit or harm in total mortality or cardiovascular events. More recently, a Women's Antioxidant Cardiovascular Study (WACS) also used a randomized factorial trial to evaluate beta-carotene, vitamin C, or vitamin E compared to a placebo [53]. The results indicated no benefit or harm for beta-carotene with a relative risk of 1.02. Another trial in women, the Women's Health Study (WHS) found no beneficial or harmful effects of beta-carotene supplementation for cancer or cardiovascular disease [54]. The first two trials involve participants who were at higher risk for lung cancer while the latter three trials involved participants at lower risk. For three of the trials [50–52], the placebo arm confirmed the observation that low serum beta-carotene was associated with higher risk of lung cancer, despite either harmful or no effects from increasing those levels when compared to a placebo arm. Clearly, serum beta-carotene as a useful biomarker for risk failed as a surrogate for clinically meaningful outcomes.

Cardiology provides two more recent examples of biomarkers, C-reactive protein (CRP) and troponin, neither of which qualified as a surrogate [2] for cardiovascular outcomes. Although mortality from heart disease has fallen, it remains the leading cause of death in the USA so research for more prevention and intervention continues. Beyond the known risk factors of lipids, blood pressure, diabetes and obesity, inflammation is now considered to have an impact on the progression of cardiovascular disease [55]. One inflammation biomarker is CRP which is low in normal individuals but increases with acute episodes such as a heart attack. CRP has some predictive ability for future coronary events. There are now standardized low-cost assays for CRP that meet criteria for analytical validation.

While CRP is correlated with cardiovascular events and thus can be used as a predictor of risk, it is not known whether it is in the causal pathway. That is, even if CRP is just tracking along with other unknown biomarkers that are on the causal pathway, it will still correlate with the clinical outcome and thus be useful for risk assessment. However, it will not be useful to monitor change in CRP since it is not on the causal pathway. Thus this biomarker of inflammation does not qualify as a surrogate outcome without further studies. One major randomized placebo-controlled trial, JUPITER, evaluated the effect of a statin, rosuvastatin, in preventing cardiovascular events in individuals with high CRP but lower or moderate values of LDL cholesterol. JUPITER showed a reduction in new cardiovascular events using this particular statin [56]. However, further analyses showed that LDL reductions and CRP reductions were weakly correlated. Thus, JUPITER was not able to demonstrate that CRP is in the causal pathway and thus not a validated surrogate. However, it may have utility as a predictor of cardiovascular risk and be used in that manner [2].

Troponin is another biomarker of interest. In fact, troponin is used as a biomarker in the current definition of a myocardial infarction or heart attack but in combination with other factors such as changes in electrocardiograms and myocardial enzyme measurements [57–59]. It is a protein that is involved with the function of both cardiac and skeletal muscle function. Subunits of troponin can be defined which are cardiac muscle specific. High levels of troponin do not automatically suggest an acute myocardial event, so it is not in the primary causal pathway. However, it is the preferred biomarker included in the definition of a heart attack and has met accepted standards of good clinical laboratory measurement. Still, the analytic validation is not complete [2]. Clinical data indicate that high levels of troponin indicate a higher risk of mortality. Thus, it passes the first requirement for surrogate qualification. However, to date there is limited evidence that suggests reducing troponin levels would improve mortality risk.

Finally, we examine the use of epogen in kidney failure patients to maintain hematocrit levels. Clinicians believe that it is important to treat anemia to maintain an adequate hematocrit level. A class of erythropoietin-stimulating agents was developed to increase hematocrit levels. One such drug, epogen, was evaluated in a randomized placebo clinical trial called TREAT (Trial to Reduce cardiovascular Events with Aranesp Therapy), which was a randomized placebo-controlled clinical trial, epogen vs. placebo plus standard of care, in type 2 diabetes patients with

kidney failure [60]. This drug was known to significantly improve hematocrit levels, considered to be a biomarker for type 2 diabetes risk and a surrogate for clinical outcome. Many investigators believed that TREAT was even unethical to start since the evidence that epogen increased hematocrit was established, and thus presumed to produce a resulting clinical benefit, breaking their equipoise between epogen and placebo[60]. Slightly over two thousand patients were randomized to each intervention. Compliance to the study medication was excellent and the resulting hemoglobin levels were statistically and clinically higher on the epogen-treated patients compared to placebo treated. Nevertheless, the composite endpoint of death, nonfatal heart attack, nonfatal stroke, heart failure, and myocardial ischemia had a hazard ratio of 1.05 (0.94–1.17), in favor of placebo. The outcome of end-stage renal disease or death had a hazard ratio of 1.06 (0.95–1.19), similar to the composite outcome. There was a significant difference in fatal and nonfatal stroke frequency, with a hazard ratio of 1.92 (1.38–2.68, $P < 0.001$) in favor of placebo. Thus, the biomarker of hemoglobin levels would not qualify as a surrogate since this trial demonstrated a dramatic effect on hemoglobin level increase with no effect on clinical outcomes, but with a significant adverse effect in fatal and nonfatal stroke. Obviously, not all of the clinical effect of epogen was captured by simply measuring hematocrit.

In early AIDS research, measures of immune response such as CD4 cell counts were used as biomarkers with the hope of being a valid surrogate. AIDS is a disease that compromises the body's immune system. One therapeutic strategy is to help the immune system recover as measured by CD4 cell counts. Low CD4 cell count was considered a predictor of mortality and morbidity in AIDS patients. Fleming and DeMets [1] report several examples where improved or positive changes in CD4 cell count did not convey clinical benefit in either mortality or morbidity. Recently, another study group has reported similar results [61]. Two separate trials evaluating two cohorts defined by their baseline CD4 counts compared interleukin-2 plus antiretroviral therapy with antiretroviral therapy alone. Despite substantial and sustained elevations of CD4 count over several years, there was no significant clinical benefit in either study.

6 Summary

The need for randomized clinical trials to evaluate new interventions will continue to be the best and primary methodology. As a result, interest in efficient trial designs will include the potential use of biomarkers as a surrogate for clinical outcomes. As described, besides being able to measure the biomarker adequately, there are two critical requirements before it can be relied on completely. Those requirements are that a reasonably strong correlation exists between the biomarker and the clinical outcome, and reasonable certainty that the biomarker is capturing all of the effects of the intervention including harmful effects. Ideally, there should be perfect correlation and 100% certainty. These stringent conditions are rarely met.

How strong the correlation must be and how certain we are about capturing all of the effects will depend on the context of the intended use. That is the point of the IOM's third criterion [2]. Thus, no specific correlation and level of certainty can be specified without considering the context of intended use. With alarming frequency, reliance on a biomarker as a surrogate for clinical outcomes has resulted in numerous interventions being utilized without appreciation of other effects, especially harmful effects. Only later, when subsequent trials were conducted were the harmful effects discovered. Biomarkers will, however, continue to play a critical role in the development of new drugs, devices, and other interventions. Some biomarkers may also be useful in identifying patient risk. However, for now the recent IOM report [2] is consistent with the recommendations of Fleming and DeMets [1] made a decade earlier that biomarkers should not be relied upon as a surrogate for clinically relevant outcomes in Phase III clinical trials.

References

1. Fleming TR, DeMets DL (1996) Surrogate endpoints in clinical trials: are we being mislead? Ann Intern Med 125:605–613
2. Committee on Qualifications of Biomarkers and Surrogate Endpoints in Chronic Disease, Michael C, Ball J (eds) (2010) Evaluation of biomarkers and surrogate endpoints in chronic disease. National Academies Press, Washington
3. Prentice RL (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 8:431–440
4. Fleming TR (1992) Evaluating therapeutic interventions: some issues and experiences (with discussion and rejoinder). Stat Sci 7:428–456
5. Fleming TR (1990) Evaluation of active control trials in AIDS. J Acquir Immun Defic Syndr 3(Suppl 2):S82–S87
6. Johnson JR, Temple R (1985) Food and Drug Administration requirements for approval of new anticancer drugs. Cancer Treat Rep 69:1155–1159
7. Ellenberg SS, Hamilton JM (1989) Surrogate endpoints in clinical trials: cancer. Stat Med 8:405–413
8. Fleming TR, Prentice RL, Pepe MS, Glidden D (1994) Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. Stat Med 3:955–968
9. Herson J (1989) The use of surrogate endpoints in clinical trials. Stat Med 8:403–404
10. Kosorok MR, Fleming TR (1993) Using surrogate failure time data to increase cost effectiveness in clinical trials. Biometrika 80:823–833
11. Machado SG, Gail MH, Ellengerg SS (1990) On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection. J Acquir Immune Defic Syndr 3:1065–1073
12. Pepe MS, Reilly M, Fleming TR (1994) Auxiliary outcome data and the mean score method. J Stat Plan Inference 42:137–160
13. Wittes J, Lakatos E, Probstfield J (1989) Surrogate endpoints in clinical trials: cardiovascular diseases. Stat Med 8:415–425
14. Ellenberg SS (1991) Surrogate end points in clinical trials [Editorial]. BMJ 302:63–64
15. Fleming TR (1994) Surrogate markers in AIDS and cancer trials. Stat Med 13:1423–1435
16. Lagakos SW, Hoth DF (1992) Surrogate markers in AIDS: where are we? Ann Intern Med 116:599–601
17. Boissel JP, Collet JP, Moleur P, Haugh M (1992) Surrogate endpoints: a basis for a rationale approach. Eur J Clin Pharmacol 43:235–244

18. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989) Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 321:406–412
19. Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH et al (1991) Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. N Engl J Med 324:781–788
20. The Cardiac Arrhythmia Suppression Trial Investigators (1992) Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. N Engl J Med 327:227–233
21. Coplen SE, Antman EM, Berlin JA, Hewitt P, Chalmers TC (1990) Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion. A meta-analysis of randomized control trials. Circulation 82:1106–1116
22. Hine LK, Laird N, Hewitt P, Chalmers TC (1989) Meta-analytic evidence against prophylactic use of lidocaine in acute myocardial infarction. Arch Intern Med 149:2694–2698
23. MacMahon S, Collins R, Peto R, Koster RW, Yusuf S (1988) Effects of prophylactic lidocaine in suspected acute myocardial infarction. An overview of results from the randomized, controlled trials. JAMA 260:1910–1916
24. Rossouw JE, Lewis B, Rifkind BM (1990) The value of lowering cholesterol after myocardial infarction. N Engl J Med 323:1112–1119
25. The Coronary Drug Project Research Group (1975) Clofibrate and niacin in coronary heart disease. JAMA 231:360–381
26. Gordon DJ (1994) Cholesterol lowering and total mortality. In: Rifkind BM (ed) Contemporary issues in cholesterol lowering: clinical and population aspects. Marcel Dekker, New York
27. (1994) Randomised trial of cholesterol lower in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 344:1383–1389
28. Fogoros RN (2001) The Baycol recall, what it means, Heart Health Center, About.com, Aug 13, 2001
29. Barter PJ, Caulfield M, Eriksson M, Grundy SM, Kastelein JJ, Komajda M, Lopez-Sendon J, Mosca L, Tardiff JC, Waters DD, Shear CL, Revkin JH, Buhr KA, Fisher MR, Tall AR, Brewer B, ILLUMINATE Investigators (2007) Effects of torcetrapib in patients at high risk for coronary events. N Engl J Med 357:2109–2122
30. Stampfer M, Colditz G (1991) Estrogen replacement therapy and coronary heart disease, a quantitative assessment of the epidemiological evidence. Prev Med 20:47–63
31. Grady D, Rueben SB, Pettiti DB et al (1992) Hormone therapy to prevent heart disease and prolong life in postmenopausal women. Ann Int Med 117:1102–1109
32. Writing Group for the Women's Health Initiative Investigators (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the Women's Health Initiative randomized controlled trial. JAMA 288(3):321–333
33. Women's Health Initiative Investigators Steering Committee (2004) Effects of conjugated equine estrogen in postmenopausal women with a hysterectomy: the Women's Health Initiative randomized clinical trial. JAMA 291:1701–1712
34. Coronary Drug Project Research Group (1973) The Coronary Drug Project: Findings leading to discontinuation of the 2.5-mg/day estrogen group. JAMA 226:652–657
35. Collins R, Peto R, MacMahon S, Hebert P, Fiebach NH, Eberlein KA et al (1990) Blood pressure, stroke and coronary heart disease. Part 2, short-term reduction in blood pressure: overview of randomized drug trials in their epidemiological context. Lancet 335:827–838
36. Hypertension Detection and Follow-up Program Cooperative Group (1979) Five-year finding of the hypertension detection and follow-up program. 1. Reduction in mortality of persons with high blood pressure: including mild hypertension. JAMA 242:2562–2571
37. Furberg CD, Berglund G, Manolio TA, Psaty BM (1994) Overtreatment and undertreatment of hypertention. J Intern Med 235:387–397
38. Psaty BM, Heckbert SR, Koepsell TD, Siscovick DS, Lemaitre R, Smith NL et al (1996) The risk of incident myocardial infarction associated with anti-hypertensive drug therapies [Abstract]. Circulation 91:925