VOLUME 28 · NUMBER 29 · OCTOBER 10 2010

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

Evaluation of Treatment-Effect Heterogeneity Using Biomarkers Measured on a Continuous Scale: Subpopulation Treatment Effect Pattern Plot

Ann A. Lazar, Bernard F. Cole, Marco Bonetti, and Richard D. Gelber

From the Dana-Farber Cancer Institute; Harvard School of Public Health; Harvard Medical School, Boston, MA; University of Vermont, Burlington, VT; University of California, San Francisco, San Francisco, CA; and Bocconi University, Milan, Italy.

Submitted January 26, 2010; accepted July 22, 2010; published online ahead of print at www.jco.org on September 13, 2010.

Supported in part by Grants No. T32 CA-09337, CA-23318, P30-DE-020752, and CA-75362 from the National Institutes of Health, National Cancer Institute, Bethesda, MD, and by the Italian Ministry of Education, University, and Research protocol 2007AYHZWC (M.B.).

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Clinical Trials repository link available on JCO.org.

Corresponding author: Richard D. Gelber, PhD, Department of Biostatistics, Harvard School of Public Health, Dana-Farber Cancer Institute, Boston, MA 02115; e-mail: gelber@jimmy.harvard.edu.

0732-183X/10/2829-4539/$20.00

DOI: 10.1200/JCO.2009.27.9182

ABSTRACT

The discovery of biomarkers that predict treatment effectiveness has great potential for improving medical care, particularly in oncology. These biomarkers are increasingly reported on a continuous scale, allowing investigators to explore how treatment efficacy varies as the biomarker values continuously increase, as opposed to using arbitrary categories of expression levels resulting in a loss of information. In the age of biomarkers as continuous predictors (eg, expression level percentage rather than positive *v* negative), alternatives to such dichotomized analyses are needed. The purpose of this article is to provide an overview of an intuitive statistical approach—the subpopulation treatment effect pattern plot (STEPP)—for evaluating treatment-effect heterogeneity when a biomarker is measured on a continuous scale. STEPP graphically explores the patterns of treatment effect across overlapping intervals of the biomarker values. As an example, STEPP methodology is used to explore patterns of treatment effect for varying levels of the biomarker Ki-67 in the BIG (Breast International Group) 1-98 randomized clinical trial comparing letrozole with tamoxifen as adjuvant therapy for postmenopausal women with hormone receptor–positive breast cancer. STEPP analyses showed patients with higher Ki-67 values who were assigned to receive tamoxifen had the poorest prognosis and may benefit most from letrozole.

J Clin Oncol 28:4539-4544. © 2010 by American Society of Clinical Oncology

INTRODUCTION

The discovery of biomarkers that predict treatment effectiveness has great potential for improving medical care, particularly in oncology.[1] Recent improvements in technology allow for the efficient ascertainment of multiple biomarkers and support the collection of quantitative information on a continuous scale. For example, in breast cancer clinical trials, the results of steroid hormone–receptor assays are increasingly reported on a continuous scale, allowing investigators to explore how treatment efficacy varies as estrogen receptor (ER) expression values continuously increase, as opposed to using arbitrary categories of expression levels.

Typical analytic approaches in cancer clinical trials evaluate treatment-effect modification, also called interaction or treatment-effect heterogeneity, by first defining (often arbitrary) patient subgroups based on biomarker expression level. Treatment comparisons are then performed within each subgroup, and the results are assessed for heterogeneity. Regression methods (eg, proportional hazards regression[2] and cumulative incidence regression[3]) are also typically used to evaluate whether biomarker status is associated with treatment efficacy.

The approach of categorizing biomarker expression may fail to fully identify the worth of the biomarker as a predictor of treatment efficacy, because categorization results in a loss of information.[4] If the biomarker is measured on a continuous scale, the analytic approach employed should ideally make use of all available information with a minimum number of assumptions. In addition, the method of analysis should have the capability of detecting a wide range of patterns of biomarker effect.

This article provides an overview of the subpopulation treatment effect pattern plot (STEPP)[5-7] method for exploring treatment-effect heterogeneity as biomarker expression varies along a continuum. STEPP methodology is designed for and has been applied to data derived from comparative clinical trials.[8-15] STEPP estimates—and displays graphically—the treatment effect along the continuous biomarker scale using overlapping patient subgroups. In standard STEPP analysis, the treatment effect is given by the absolute difference between the treatment group and control group survival curves at a specified time point (eg, 5-year survival rate). To expand the capability of STEPP, we include two other useful measures of treatment efficacy: hazard ratios and cumulative incidence estimates. The use

of hazard ratios obviates the need to specify a time point for treatment comparison. Cumulative incidence methods allow STEPP to be used when competing causes of failure are relevant (eg, when analyzing local failure and treating other failure types as competing risks). The primary advantage of STEPP over other methods is that STEPP can detect nonlinear patterns of treatment-effect heterogeneity while making few or no distributional or parametric assumptions. We illustrate STEPP with an analysis of the BIG (Breast International Group) 1-98 clinical trial,[12,16-18] which compared adjuvant letrozole with tamoxifen in the treatment of postmenopausal women with hormone receptor–positive breast cancer.

STEPP METHOD

STEPP Basics

STEPP methodology[5-7] examines interaction between treatment and covariate by estimating the treatment effect within overlapping subpopulations of patients, where the subpopulations correspond to values of the covariate along its continuum. The overlapping subpopulations are constructed as follows: Patients are ordered from lowest to highest value of the covariate. The investigator chooses two quantities: r_1 and r_2, where r_1 is the largest number of patients in common (or overlapping) among consecutive subpopulations, and r_2 is the number of patients in each subpopulation ($r_2 > r_1$). The first subpopulation consists of the r_2 patients with the lowest covariate values. The next subpopulation is formed by removing r_2 minus r_1 patients with the lowest covariate values from the current subpopulation and replacing them with the next r_2 minus r_1 patients in the ordered list. This process continues until all patients have been included in at least one subpopulation. Note that patients can contribute to several subpopulations. This approach is called sliding-window STEPP.[5-7]

After the overlapping subpopulations are identified, the treatment effect is estimated within each subpopulation using a standard approach, such as the Kaplan-Meier[19] product limit method, with the treatment effect represented by the absolute difference between two survival curves at a particular time point. STEPP analysis results are then shown graphically. STEPP analysis can be performed using the R software package (R Foundation for Statistical Computing, Vienna, Austria; https://sites.google.com/site/stepprpackage).

STEPP Extensions

STEPP can be implemented using hazard ratios to describe the treatment effect. The hazard ratio can be obtained from observed minus expected (*O*-E) numbers of events in a fashion similar to that used to compute the log-rank statistic for comparing survival curves. If *O*-E represents the log-rank statistic, and V represents its variance, an estimate of the hazard ratio can be obtained by exponentiation of [(*O*-E)/V].[20] This estimate requires no assumptions about the underlying distribution of survival times.[21] It also obviates the need to specify a time point for comparing survival curves.

Treatment effects on disease-specific end points can also be estimated using methodology for competing risk analysis.[3,22-24] Focusing on disease-specific events in the competing risk setting provides the most direct connection between the biomarker and its relationship to treatment and outcome.[25] Therefore, we introduce STEPP for comparing treatment groups with respect to disease-specific cumulative incidence estimated using the standard method for competing risk analysis.

Statistical significance in a STEPP analysis is calculated using a permutation test.[7] A two-sided *P* value less than .05 indicates significant treatment-effect heterogeneity.

APPLICATION: KI-67 AND LETROZOLE EFFECTIVENESS IN BREAST CANCER

A well-known predictor of breast cancer prognosis is the tumor proliferation fraction,[26] which is associated with the degree of effectiveness of chemotherapy. Ki-67, a nuclear protein present in cycling cells, is an indicator of tumor proliferation.[27] High Ki-67 labeling index (LI) is associated with a strong response to preoperative chemotherapy.[28,29] During preoperative endocrine therapy, decline in this biomarker is linked to pathologic tumor response.[30]

The prognostic and predictive value of Ki-67 LI were evaluated in the BIG 1-98 study,[9] an international, double-blind phase III clinical trial of 8,010 postmenopausal women with early stage invasive breast cancer, who were randomly assigned to one of four adjuvant endocrine therapy arms: letrozole, tamoxifen, or sequences of these agents (letrozole to tamoxifen, tamoxifen to letrozole).[9] The primary trial end point was disease-free survival (DFS), defined as the length of time from randomization to the first event of invasive recurrence in local, regional, or distant sites; a new invasive breast cancer in the contralateral breast; any second nonbreast malignancy; or death as a result of any cause. Previous trial reports that compared the two monotherapy arms demonstrated significant DFS improvement in patients initially assigned to letrozole compared with tamoxifen.[16,18]

Analysis of Absolute Treatment Effects

STEPP analysis of 4-year DFS was used to explore the patterns of treatment effectiveness across the continuum of Ki-67 LI percentages (Ki-67 LI range, 0% to 90%). Of the 4,922 patients who were randomly assigned to receive 5 years of monotherapy with either letrozole or tamoxifen, 2,685 patients had tumors with centrally confirmed ER expression and tumor material available for Ki-67 LI determination in the central laboratory, as described by Viale et al.[9] The database was the same one used by Viale et al, with a median follow-up of 51 months. The 4-year time point was selected to coincide with the time point used in previous analyses of BIG 1-98 data.

To construct the overlapping subpopulations for the STEPP analysis, we set $r_2 = 150$ patients as the size of each subpopulation, and we set $r_1 = 50$ as the number of patients included within consecutive overlapping subpopulations. Figure 1A summarizes the 4-year DFS percentage for letrozole versus tamoxifen as Ki-67 LI increases. The figure suggests that higher Ki-67 was associated with lower 4-year DFS percentages, especially for tamoxifen. Subpopulations with high Ki-67 LI had the greatest magnitude of treatment difference, indicating benefit for letrozole compared with tamoxifen. Figure 1B shows the difference in 4-year DFS percentages (letrozole minus tamoxifen; differences > zero favor letrozole) and 95% point-wise CIs. For patients with the highest Ki-67 LI levels (ie, those in the subpopulation with median Ki-67 LI equal to 47%), the estimated absolute difference in 4-year DFS was nearly 35% in favor of letrozole. Thus, STEPP analysis provided evidence of heterogeneous treatment effects related to the value of Ki-67 LI ($P = .03$ for interaction; Fig 1B).

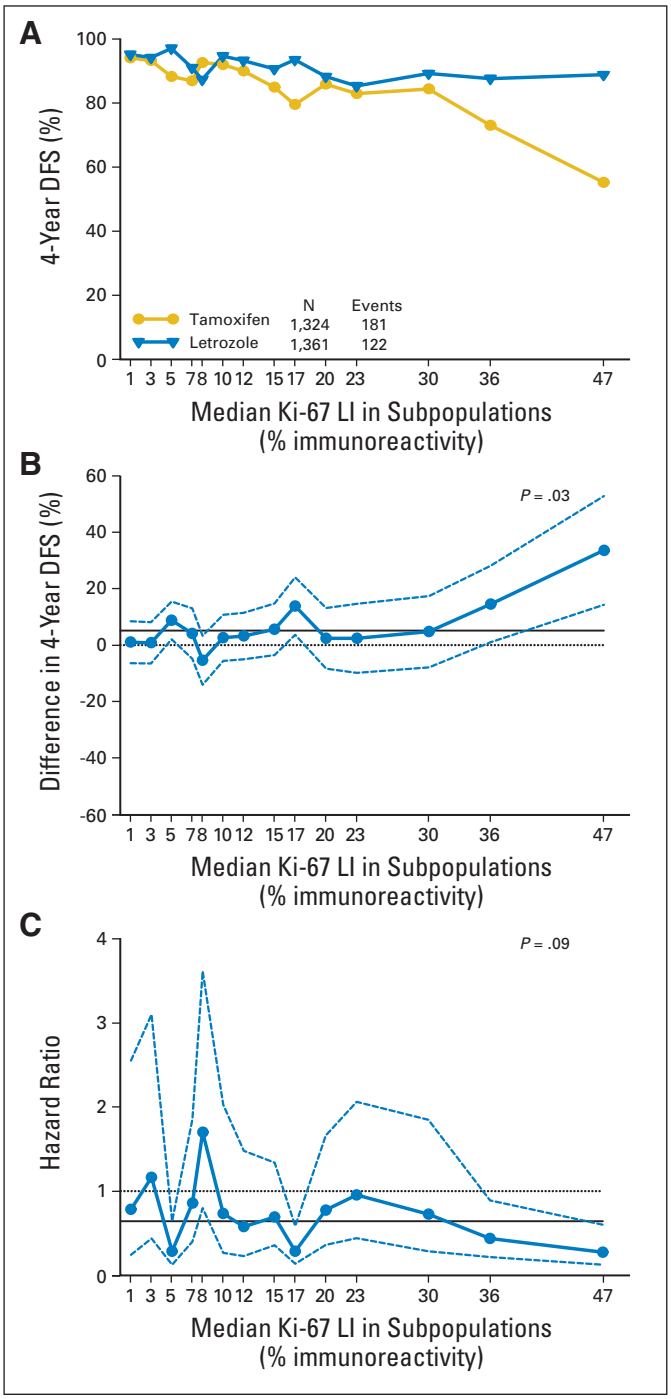

Fig 1. Subpopulation treatment effect pattern plot analysis of the treatment effect of letrozole *v* tamoxifen as measured by (A) 4-year disease-free survival (DFS), (B) difference in 4-year DFS (letrozole minus tamoxifen; > zero suggested letrozole better; otherwise, tamoxifen better), and (C) hazard ratio (letrozole *v* tamoxifen; < one suggested letrozole better; otherwise, tamoxifen better) with corresponding 95% point-wise CIs (dashed blue lines). The x-axes indicate median percentage of Ki-67 labeling index (LI) for patients in each of the overlapping subpopulations. Each subpopulation contains approximately 150 (r_2) patients and approximately 50 (r_1) overlapping patients. Solid black lines indicate overall treatment effect, and dotted black lines indicate no effect. *P* values are from interaction test.

Analysis of Relative Treatment Effects

We also used STEPP to explore patterns of relative treatment effectiveness based on hazard ratios across patient subpopulations. For high Ki-67 LI percentages, the estimated hazard ratio of a DFS event was lower for patients receiving letrozole than for those receiving tamoxifen (Fig 1C; a hazard ratio < one indicates that letrozole was better than tamoxifen). For the subpopulation with the highest Ki-67 LI, the hazard ratio of a DFS event for patients in the letrozole group was less than half that for patients in the tamoxifen group. STEPP analysis results were suggestive of heterogeneous relative treatment effects, although this was not statistically significant ($P = .09$ for interaction; Fig 1C).

As an alternative to the STEPP approach, we used Cox proportional hazards modeling to evaluate a treatment by covariate interaction effect on DFS. Three models were considered to evaluate different forms of the covariate. First, we used the median cutoff from Ki-67 LI distribution, where levels of the biomarker covariate Ki-67 LI were dichotomized as high (> 10%) or low (≤ 10%). The treatment by covariate interaction was not statistically significant ($P = .11$). We then used quartiles of the Ki-67 LI distribution to define patient subgroups as follows: high (19% to 90%), medium high (11% to 18%), low medium (6% to 10%), and low (0% to 5% [reference category]). The interaction test did not provide statistically significant results ($P = .10$). Finally, we used Ki-67 LI percentage as a continuous covariate in the Cox model. In this case, the treatment by Ki-67 LI interaction was borderline statistically significant ($P = .05$). Overall, although these analyses all suggested the presence of treatment-effect heterogeneity in terms of hazard ratios, no statistically significant heterogeneity was detected.

Competing Risk Analysis: Absolute Treatment Effects

STEPP analysis in the competing risk setting is illustrated in Figure 2. The end point was 4-year cumulative incidence of breast cancer relapse (in local, regional, or distant sites) in the presence of competing risks of second (nonbreast) primaries or death before breast cancer relapse. The cumulative incidence of breast cancer relapse increased with increasing Ki-67 values for both treatment groups (Fig 2A). The treatment curves separate for subpopulations with median Ki-67 LI values of 15% or more (Fig 2A), and the magnitude of this difference in favor of letrozole peaked for high Ki-67 LI (Fig 2B). STEPP analysis provided evidence of heterogeneous treatment effects related to the value of Ki-67 LI percentages ($P = .02$ for interaction; Fig 2B).

Competing Risk Analysis: Relative Treatment Effects

The estimated hazard ratio for a breast cancer relapse event tended to be less than 1.0, suggesting an overall advantage for letrozole relative to tamoxifen. However, no significant treatment-effect heterogeneity was found as Ki-67 LI ranged from low to high ($P = .35$ for interaction; Fig 2C). These findings were consistent with those from a competing risk regression analysis, which we also applied to the data, defining patient subgroups based on Ki-67 LI expression (two groups using median cutoff and four groups using quartiles).

DISCUSSION

Technologic advancements provide the ability to measure an increasing number of biomarkers on a continuous scale. Modern large-scale

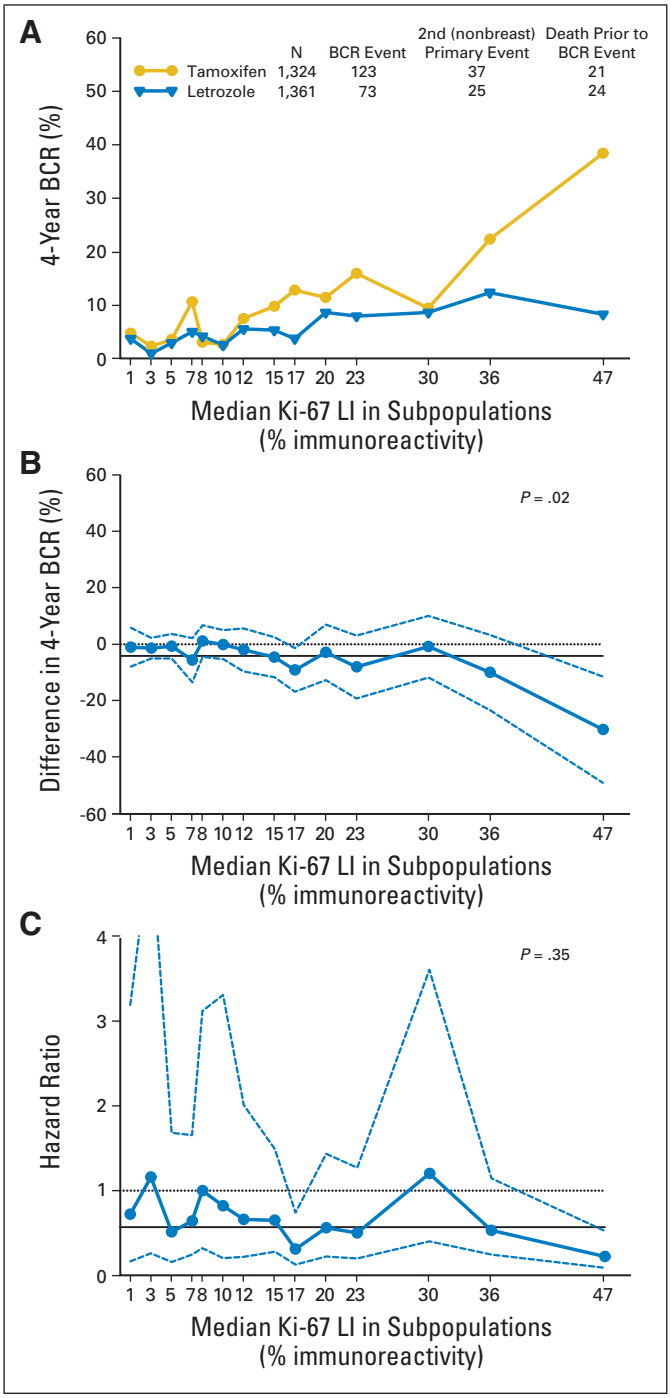

Fig 2. Subpopulation treatment effect pattern plot analysis of the treatment effect of tamoxifen *v* letrozole as measured by (A) 4-year cumulative incidence of breast cancer recurrence (BCR), (B) difference in 4-year cumulative incidence of BCR (letrozole minus tamoxifen; < zero suggested letrozole better; otherwise, tamoxifen better), and (C) hazard ratio (letrozole *v* tamoxifen; < one suggested letrozole better; otherwise, tamoxifen better) with corresponding 95% point-wise CIs (dashed blue lines). The x-axes indicate median percentage of Ki-67 labeling index (LI) for patients in each of the overlapping subpopulations. Each subpopulation contains approximately 150 (r_2) patients and approximately 50 (r_1) overlapping patients. Solid black lines indicate overall treatment effect, and dotted black lines indicate no effect. *P* values are from interaction test.

cancer clinical trials include the quantitative assessment of selected biomarkers thought to be associated with clinical outcome and treatment effectiveness. This approach is meant to aid in the identification of subgroups of patients most likely to benefit from a particular treatment modality. Improved analytic techniques are needed to assess potentially complex associations between biomarker expression level and treatment effect. These associations may not follow a linear pattern or may not be detected with standard approaches, such as those that categorize patients according to biomarker level (eg, those based on quartiles).[4] In this article, we provide an overview of STEPP methodology for evaluating treatment-effect heterogeneity over the range of values for a continuous covariate, and we illustrate the approach in a breast cancer clinical trial.

Using STEPP analyses, we explored the patterns of treatment-effect heterogeneity for a breast cancer biomarker—Ki-67—in postmenopausal women enrolled onto the BIG 1-98 study. The treatment effectiveness patterns suggested that patients with higher Ki-67 LI percentages who were assigned to receive tamoxifen had poorer prognosis, and these patients may benefit most from letrozole treatment compared with tamoxifen. This benefit may be explained by the reduced residual circulating estrogen levels in patients receiving an aromatase inhibitor such as letrozole. Elevated levels of residual estrogen combined with high levels of growth factor receptors may be responsible for worsening prognosis for women receiving tamoxifen because of the activated membrane ER.[31] Despite the biologic plausibility of the observed results, examining the role of Ki-67 LI as a predictive factor was not an analysis specified in the BIG 1-98 protocol written in 1998, and thus the results should be interpreted cautiously.

In general, retrospective subgroup analyses are associated with many well-documented problems.[4,32-34] However, as stated by Lagakos,[33] "avoiding any presentation of subgroup analyses because of their history of being overinterpreted is a steep price to pay for a problem that can be remedied by more responsible analysis and reporting." Recent guidelines[34] have been proposed to reduce the risk of overinterpretation of results from subgroup analysis. Such guidelines are appropriate to apply to STEPP analysis as well.

One limitation of STEPP is the need to specify the number of patients per subpopulation (r_2) and the number to be exchanged to form subsequent subpopulations (r_2 minus r_1). The estimation of interaction effects will vary for different r_1 and r_2. Therefore, we recommend using a variety of r_1 and r_2 values to assess the stability of the results. This approach in our BIG 1-98 example helped us identify the subgroups driving the significant *P* value, which we consistently found to be the patients with the highest Ki-67 LI percentages. Finally, we recommend evaluating a variety of time points, such as 3- and 4-year DFS. In the BIG 1-98 example, the results were consistent across different time points (2-, 3-, or 4-year DFS).

The type of end point selected will also influence the interpretation of STEPP analysis results and potentially the impact of the results on clinical practice.[35] For example, an interaction detected between a covariate and treatment effect measured on the absolute scale (eg, 4-year cumulative incidence of breast cancer relapse) may not be detected if the treatment effect is measured on the relative scale (eg, hazard ratio).[36] These different relationships were evident in the BIG 1-98 evaluation, especially in the competing risk setting. Ki-67 LI is a prognostic factor such that even if the relative treatment effects do not vary widely across Ki-67 LI subpopulations, the absolute effect will be larger for the cohorts at higher risk for relapse. Regardless of the choice

of end point, it is also relevant to note that tests for interaction are in general underpowered.

Importantly, STEPP is useful for evaluating treatment-effect heterogeneity in both absolute (eg, absolute difference between two survival curves at a particular time point) and relative terms (eg, hazard ratio). Relative effects are primarily useful for measuring treatment effectiveness relative to a control group in a general population of patients. Absolute effects are clinically useful for treatment decision making in individual patients, where the absolute benefit must be weighed against any risks associated with a particular therapy.

Other statistical approaches can be used to evaluate interaction between treatment and covariate for cancer-related biomarkers. One approach is multivariable fractional polynomial interaction (MFPI), which can also produce treatment effect plots.[37,38] MFPI relies on regression modeling via searching through several covariates and forms of treatment by covariate interactions. This is accomplished through a proportional hazards model. Another approach combines STEPP with a smoothing technique known as locally weighted regression,[39] which is useful for studying the association between a biomarker measured on a continuous scale and treatment effectiveness. The martingale residual plot with locally weighted scatterplot smoothing is an approach useful for examining the relationship between the end point and continuous covariate.[40] Splines can also be considered, including regression and smoothing splines.[40] Splines as well as the martingale residual plot may provide alternatives to evaluation of treatment by covariate interaction.

When testing for heterogeneity of treatment effects, we recommend applying multiple analytic approaches to assess the stability of results and conclusions. For example, we also applied MFPI methodology to the BIG 1-98 study and obtained results similar to those seen with STEPP (Fig 1C). When different analytic approaches yield inconsistent results, the preference of one approach over another will hinge on the appropriateness of the assumptions made in the analysis. In this case, we recommend reporting results from all analyses along with a rationale for preference for a particular approach. Lack of consistency may be a result of sparseness of data, which would need to be addressed by enriching the data source or performing additional research and data gathering.

STEPP analysis offers several advantages compared with other statistical approaches. STEPP does not require predefinition of specific cutoff points for developing patient subgroups. Treatment-effect heterogeneity is illustrated graphically, allowing for a convenient exploratory evaluation. STEPP does not rely on the appropriateness of a regression model. Estimates derived from STEPP analysis can be presented with CIs or confidence bands. STEPP provides an overall *P* value for testing whether treatment-effect heterogeneity is significant. STEPP analysis can be based on absolute or relative treatment effects. Finally, STEPP can be applied in the presence of competing risks. This advantage is particularly relevant to the study of biomarkers because of the potential for a biomarker to be most strongly associated with a particular type of clinical outcome (eg, local recurrence).[41]

Discovery of potentially important biomarkers is facilitated by the availability of modern statistical tools for evaluating treatment covariate interactions, such as STEPP. These statistical approaches enable clinicians to identify biomarker candidates. The complex roles biomarkers play in differential treatment effectiveness will be determined after extensive validation studies, as replication and confirmation are hallmarks of science.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Although all authors completed the disclosure declaration, the following author(s) indicated a financial or other interest that is relevant to the subject matter under consideration in this article. Certain relationships marked with a "U" are those for which no compensation was received; those relationships marked with a "C" were compensated. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Employment or Leadership Position: None **Consultant or Advisory Role:** None **Stock Ownership:** None **Honoraria:** None **Research Funding:** Richard D. Gelber, Novartis **Expert Testimony:** None **Other Remuneration:** None

AUTHOR CONTRIBUTIONS

Conception and design: Ann A. Lazar, Bernard F. Cole, Marco Bonetti, Richard D. Gelber
Financial support: Richard D. Gelber
Data analysis and interpretation: Ann A. Lazar, Bernard F. Cole, Marco Bonetti
Manuscript writing: Ann A. Lazar, Bernard F. Cole, Richard D. Gelber
Final approval of manuscript: Ann A. Lazar, Bernard F. Cole, Marco Bonetti, Richard D. Gelber

REFERENCES

1. Yeatman TJ: Predictive biomarkers: Identification and verification. J Clin Oncol 27:2743-2744, 2009
2. Cox DR: Regression models and life tables (with discussion). J Roy Stat Soc B 34:187-220, 1972
3. Fine JP, Gray RJ: A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94:496-509, 1999
4. Royston P, Altman D, Sauerbrei W: Dichotomizing continuous predictors in multiple regression: A bad idea. Stat Med 25:127-141, 2006
5. Bonetti M, Gelber RD: A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. Stat Med 19: 2595-2609, 2000
6. Bonetti M, Gelber RD: Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics 5:465-481, 2004
7. Bonetti M, Zahrieh D, Cole BF, et al: A small sample study of the STEPP approach to assessing treatment-covariate interactions in survival data. Stat Med 28:1255-1268, 2009
8. International Breast Cancer Study Group: Endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node negative breast cancer: A randomized trial. J Natl Cancer Inst 94: 1054-1055, 2002
9. Viale G, Giobbie-Hurder A, Regan MM, et al: Prognostic and predictive value of centrally reviewed Ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: Results from Breast International Group Trial 1-98 comparing adjuvant tamoxifen with letrozole. J Clin Oncol 26:5569-5575, 2008
10. Penault-Llorca F, André F, Sagain C, et al: Ki67 expression and docetaxel efficacy in patients with estrogen receptor-positive breast cancer. J Clin Oncol 27:2809-2815, 2009
11. Crivellari D, Sun Z, Coates A, et al: Letrozole compared with tamoxifen for elderly patients with endocrine-responsive early breast cancer: The BIG 1-98 trial. J Clin Oncol 26:1972-1979, 2008
12. Viale G, Regan MM, Maiorano E, et al: Chemoendocrine compared with endocrine adjuvant therapies for node-negative breast cancer: Predictive value of centrally reviewed expression of estrogen and progesterone receptors—International Breast Cancer Study Group. J Clin Oncol 28:1404-1410, 2008
13. Rasmussen BB, Regan MM, Lykkesfeldt AE, et al: Adjuvant letrozole versus tamoxifen according to centrally-assessed ERBB2 status for postmenopausal women with endocrine-responsive early

breast cancer: Supplementary results from the BIG 1-98 randomised trial. Lancet Oncol 9:23-28, 2008

14. Regan MM, Gelber RD: Predicting response to systemic treatments: Learning from the past to plan for the future. Breast 14:582-593, 2005

15. International Breast Cancer Study Group, Castiglione-Gertsch M, O'Neill A, et al: Adjuvant chemotherapy followed by goserelin versus either modality alone for premenopausal lymph node-negative breast cancer: A randomized trial. J Natl Cancer Inst 95:1833-1846, 2003

16. Breast International Group (BIG) 1-98 Collaborative Group, Thürlimann B, Keshaviah A, et al: A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. N Engl J Med 353:2747-2757, 2005

17. BIG 1-98 Collaborative Group, Mouridsen H, Giobbie-Hurder A, et al: Letrozole therapy alone or in sequence with tamoxifen in women with breast cancer. N Engl J Med 361:766-776, 2009

18. Coates AS, Keshaviah A, Thürlimann B, et al: Five years of letrozole compared with tamoxifen as initial adjuvant therapy for postmenopausal women with endocrine-responsive early breast cancer: Update of study BIG 1-98. J Clin Oncol 25:486-492, 2007

19. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457-481, 1958

20. Peto R, Pike MC, Armitage P, et al: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Br J Cancer 35:1-39, 1977

21. Early Breast Cancer Trialists' Collaborative Group: Treatment of Early Breast Cancer, Volume 1: Worldwide Evidence 1985-1990. Oxford, United Kingdom, Oxford University Press, 1990, pp 6-18

22. Kalbfleisch JD, Prentice RL: The Statistical Analysis of Failure Time Data. New York, NY, Wiley, 1980, pp 168-169

23. Gray RJ: A class of K-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat 16:1141-1154, 1988

24. Dignam JJ, Kocherginsky MN: Choice and interpretation of statistical tests used when competing risks are present. J Clin Oncol 26:4027-4034, 2008

25. Hudis CA, Barlow WE, Costantino JP, et al: Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: The STEEP system. J Clin Oncol 25:2127-2132, 2007

26. Clahsen PC, van de Velde CJ, Duval C, et al: The utility of mitotic index, oestrogen receptor and Ki-67 measurements in the creation of novel prognostic indices for node-negative breast cancer. Eur J Surg Oncol 25:356-363, 1999

27. Gerdes J, Schwab U, Lemke H, et al: Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. Int J Cancer 31:13-20, 1983

28. Chang J, Ormerod M, Powles TJ, et al: Apoptosis and proliferation as predictors of chemotherapy response in patients with breast carcinoma. Cancer 89:2145-2152, 2000

29. Archer CD, Parton M, Smith IE, et al: Early changes in apoptosis and proliferation following primary chemotherapy for breast carcinoma. Br J Cancer 89:1035-1041, 2003

30. Miller WR, White S, Dixon JM, et al: Proliferation, steroid receptors and clinical/pathological response in breast cancer treated with letrozole. Br J Cancer 94:1051-1056, 2006

31. Lewis-Wambi JS, Jordan VC: Treatment of postmenopausal breast cancer with selective estrogen receptor modulators (SERMs). Breast Dis 24: 93-105, 2005

32. Cuzick J: Forest plots and the interpretation of subgroups. Lancet 1308, 2005

33. Lagakos SW: The challenge of subgroup analyses: Reporting without distorting. N Engl J Med 354:1667-1669, 2006

34. Wang R, Lagakos SW, Ware JH, et al: Statistics in medicine: Reporting of subgroup analyses in clinical trials. N Engl J Med 357:2189-2194, 2007

35. Berrington de Gonzalez A, Cox DR: Interpretation of interaction: A review. Ann Appl Stat 1:371-385, 2007

36. Pocock S: More on subgroup analysis in clinical trials. N Engl J Med 358:2076-2077, 2008

37. Royston P, Altman D: Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Appl Stat 43:429-467, 1994

38. Royston P, Sauerbrei W: A new approach to modeling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 19:2509-2525, 2004

39. Jeong JH, Constantino JP: Application of smoothing methods to evaluate treatment-prognostic factor interactions in breast cancer data. Cancer Invest 24:288-293, 2006

40. Therneau TM, Grambsch PM: Modeling Survival Data: Extending the Cox Model. New York, NY, Springer, 2000, pp 87-126

41. Cuzick J, Sasieni P, Howell A: Should aromatase inhibitors be used as initial adjuvant treatment or sequenced after tamoxifen? Br J Cancer 94:460-464, 2006

