Graphical Presentation of Patient-Treatment Interaction Elucidated by Continuous Biomarkers*

Current Practice and Scope for Improvement

Yu-Ming Shen; Lien D. Le; Rory Wilson; Ulrich Mansmann

Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilian University Munich, Munich, Germany

Keywords

Graphical presentation, patient-treatment interaction, continuous biomarker, randomized parallel-group controlled trial

Summary

Background: Biomarkers providing evidence for patient-treatment interaction are key in the development and practice of personalized medicine. Knowledge that a patient with a specific feature – as demonstrated through a biomarker – would have an advantage under a given treatment vs. a competing treatment can aid immensely in medical decision-making. Statistical strategies to establish evidence of continuous biomarkers are complex and their formal results are thus not easy to communicate. Good graphical representations would help to translate such findings for use in the clinical community. Although general guidelines on how to present figures in clinical reports are available, there remains little guidance for figures elucidating the role of *c*ontinuous *b*iomarkers in *p*atient-*t*reatment *i*nteraction (CBPTI).

Objectives: To combat the current lack of comprehensive reviews or adequate guides on graphical presentation within this topic, our study proposes presentation principles for CBPTI plots. In order to understand current practice, we review the development of CBPTI methodology and how CBPTI plots are currently used in clinical research.

Methods: The quality of a CBPTI plot is determined by how well the presentation provides key information for clinical decision-making. Several criteria for a good CBPTI plot are proposed, including general principles of visual display, use of units presenting absolute outcome measures, appropriate quantification of statistical uncertainty, correct display of benchmarks, and informative content for answering clinical questions especially on the quantitative advantage for an individual patient with regard to a specific treatment. We examined the development of CBPTI methodology from the years 2000–2014, and reviewed how CBPTI plots were currently used in clinical research in six major clinical journals from 2013–2014 using the principle of *theoretical saturation*. Each CBPTI plot found was assessed for appropriateness of its presentation and clinical utility.

Results: In our review, a total of seven methodological papers and five clinical reports used CBPTI plots which we categorized into four types: those that distinguish the outcome effect for each treatment group; those that show the outcome differences between treatment groups (by either partitioning all individuals into subpopulations or modelling the functional form of the interaction); those that evaluate the proportion of population impact of the biomarker; and those that show the classification accuracy of the biomarker. The current practice of utilizing CBPTI plots in clinical reports suffers from methodological shortcomings: the lack of presentation of statistical uncertainty, the outcome measure scaled by relative unit instead of absolute unit, incorrect use of benchmarks, and being non-informative in answering clinical questions.

Conclusions: There is considerable scope for improvement in the graphical representation of CBPTI in clinical reports. The current challenge is to develop instruments for high-quality graphical plots which not only convey quantitative concepts to readers with limited statistical knowledge, but also facilitate medical decision-making.

Correspondence to:
Ulrich Mansmann
Institute of Medical Informatics, Biometry and Epidemiology
Ludwig Maximilian University Munich
Marchioninistr. 15
81377 Munich
Germany
E-mail: mansmann@ibe.med.uni-muenchen.de

Methods Inf Med 2017; 55: 13–27
https://doi.org/10.3414/ME16-01-0019
received: February 17, 2016
accepted: July 14, 2016
epub ahead of print: October 26, 2016

* Supplementary material published on our website https://doi.org/10.3414/ME16-01-0019

1. Background

Consider a trial in which individuals are randomized to either standard or experimental treatment. The primary aim of such a trial is usually the estimation of the overall treatment effect. Given a weak overall effect and the effort and cost involved in the trial, the investigators are frequently motivated to search for a subgroup of patients with a reasonable response to the new treatment, often through the use of distinguishing (predictive) biomarkers. Biomarkers are thus a common tool for exploring population heterogeneity with respect to treatment response. For example, a clinically established treatment-selection biomarker is presence of the K-RAS wild type gene for selecting ce-

tuximab as the treatment for metastatic colorectal cancer patients [1].

To present treatment-biomarker interaction, graphical methods are often used. For a set of dichotomous/categorical treatment-selection biomarkers, a modified forest plot can present heterogeneity of treatment effects within subpopulations [2]. In the case of continuous biomarkers, this tool cannot be applied, except if the continuous biomarker is categorized. However, categorization of continuous variables for use as biomarkers can destroy information [3], affect randomization [4] and create statistical multiplicity issues due to selection of an optimal cut-off value [5]. Further, there can be instability of the statistical significance of the treatment-biomarker interaction depending on the number and positions of cut-off values [6]. Therefore, investigation of a continuous biomarker should initially be performed without categorization. One must make careful specification of the functional form of the relationship between the continuous biomarker and the treatment effect (either linear or nonlinear), since misspecification of the relationship can lead to loss of power and faulty interpretation [6]. Subsequent steps may involve the categorization of a continuous biomarker after careful sensitivity analyses of a variety of cut-off points.

Tools to graphically present differential effects between a continuous biomarker and specific treatments exist in the literature, but have not been developed systematically [7]. The most popular approaches are treatment-effect plots [6] and subpopulation treatment-effect pattern plots (STEPP) [8]. A treatment-effect plot describes how the treatment effect changes continuously with a biomarker by using varying coefficient models based on fractional polynomials [6]. For survival outcomes the hazard ratio (HR) is displayed on the y-axis and the range of the continuous biomarker on the x-axis. Alternatively, Bonetti et al [8] propose STEPP, which uses spline functions for exploring treatment-effect heterogeneity across the range of a continuous biomarker. Their approach is based on splitting the individuals into subgroups with respect to the biomarker of interest, and calculating the effect measures separately for each subpopulation. Subpopulations are allowed to overlap in order to increase the number of subjects contributing to each point estimate, hence increasing the precision of the individual estimates. A heterogeneous treatment effect is apparent if the effect estimates do not form a horizontal line across the continuous biomarker value. Although they formulate their method in terms of hazard ratios or differences in survival probabilities, it is straightforward to generalize this approach to other effect measures like the odds-ratio (OR) or difference (Δ).

For example, ▶Figure 1 presents a heterogeneous treatment effect for breast cancer patients undergoing tamoxifen plus chemotherapy vs. tamoxifen alone [8]. It is of interest to identify patients who have an advantage under the combination treatment by using the biomarker estrogen receptor (ER) expression. ER expression is on the x-axis, ranging from 1 fmol/mg to 660 fmol/mg. The y-axis presents the hazard ratio of the combination therapy vs the monotherapy. A value below one indicates longer disease-free survival (DFS) under the combination, a value above 1 indicates longer DFS under the monotherapy. Therefore, a benchmark line parallel to the x-axis at y = 1 is introduced. Additionally, a second dashed line parallel to the x-axis is introduced, representing the treatment effect for the overall sample in the trial. The dotted black line represents the hazard ratio for individuals with various values of the biomarker. In addition to the dotted black line, broken lines above and below indicate confidence bands.

How helpful is this plot for the clinician? Is there convincing evidence for biomarker-treatment interaction or treatment effect heterogeneity in the trial population? Is the qualitative information provided helpful for the clinician? Does this plot offer clear quantitative information on the probability to profit from the combination therapy? Can the clinician determine a subgroup of patients who profit from the combination therapy by measuring the ER expression and quantify the advantage? Can the graph help to derive individual therapeutic decisions? What are the principles for constructing a continuous biomarker in patient-treatment interaction (CBPTI) plot which is helpful in answering these clinical questions?

CBPTI results are often derived using complicated statistical algorithms. Therefore, a good graphical presentation is crucial for the communication of these complex medical research findings. Many authors have discussed strategies for graphical displays in clinical trial reports, regarding the choice of figures, styles of presentation, labelling, and their specific content [9–11]. However, although general principles are available on what constitutes good practice in developing figures, there is relatively little guidance on using graphical methods to aid in the presentation of treatment-biomarker interaction in trial reports, despite a massive amount of effort being devoted to discovering treatment-selection biomarkers. Previous papers have proposed several important aspects for good plots: absolute versus relative unit for expressing the results of trials [12, 13], the types of confidence inter-

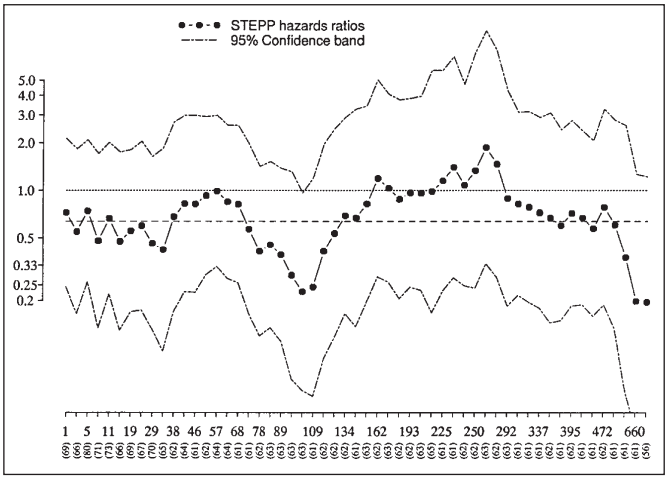

Figure 1
STEPP (sliding-window analysis) for IBCSG Trial VII data according to ER values ($n_1 = 55$, $n_2 = 60$). (Bonetti et al. Stat Med. Oct 15, 2000, p. 2601) [8]

vals/bands to quantify statistical uncertainty [14], improving direct interpretation by adding a benchmark line [2], using the same scale to facilitate comparison of candidate biomarkers [15, 16], and focusing on answering key clinical questions about the proportion of impacted patients given the use of various biomarker measures to select treatment [16, 17]. Thus, a good graphical presentation of treatment-selection biomarkers must incorporate the above elements and serve as a tool for clinical treatment decisions. The example presented in ▶Figure 1 may fulfill these criteria. But can it help to answer the above-mentioned key clinical questions?

2. Objectives

It is of key clinical interest to determine whether a specific patient characteristics (biomarker) can provide sufficient evidence to choose between two treatment options; and whether such an advantage can be quantified in easy-to-interpret measures. The quantification of an advantage in treatment outcome is necessary for a fair contrast to the potential risks of additional complications. The objective of our study is to formulate principles for CBPTI plots based on these concepts. In order to understand the current practice, we review the development of CBPTI methodology and how CBPTI plots are currently used in clinical research (it remains unclear to what extent the plots are used in clinical practice). We critically appraise each CBPTI plot and provide objective evidence as to the quality of CPBTI plots in current practice. We also add two new types of CBPTI plots and provide an R package which applies our ideas in a very simple setting, a biomarker with a linear biomarker-treatment effect relationship.

3. Methods

3.1 Criteria for a Good CBPTI Plot

A CBPTI plot should be informative for medical decision-making, i.e. it should detail a quantified advantage for choosing treatment option A versus option B for an individual patient. What are the characteristics of a CBPTI plot that ensure informative content for medical decision-making? Often CBPTI plots simply present the treatment effect across the range of the biomarker, together with the p-value of an interaction test indicating whether there is heterogeneity of treatment effect across the biomarker values. But such a presentation is insufficient for clinical application, as noted when considering the following crucial elements.

3.1.1 Absolute versus Relative Scale

The type of unit of outcome measures selected will influence the interpretation of CBPTI plots. Take as an example the BIG 1–98 trial [18]. The authors were trying to evaluate whether the Ki-67 protein could be used as a biomarker in selecting letrozole treatment in postmenopausal women with early invasive breast cancer. The results show that there is a heterogeneous treatment effect measure detected on the absolute scale (e.g., difference in 4-year disease-free survival rate) but not on the relative scale (e.g., HR). Based on the findings, how do we form a clear conclusion?

Whether absolute scale or relative scale should be used in clinical reports is still undecided. For a CPBTI plot, Rothwell et al suggested using absolute scales to detect heterogeneity of treatment among subgroups [12]. Their formation of subgroups is based on baseline risk scores estimated by specific prognostic factors in risk models. The heterogeneity of treatment effect is determined using individuals with similar risk. In contrast, Sun et al proposed the use of relative scales in subgroup analyses since relative treatment effect is constant across individuals with varying baseline risk [13]. An example of statin therapy reducing the risk of major coronary events is given in their report [13]. A meta-analysis shows that statin therapy could reduce the relative risk of major coronary events by 29.2%. If we consider using absolute risk reduction among patients with varying baseline risk, an evident heterogeneous treatment effect would exist when comparing a low baseline risk patient (1.5%, from 5% to 3.5%) with a high baseline risk patient (14.6%, from 50% to 35.4%). Therefore, given the known prognostic factors that allow the definition of subgroups, if there is no heterogeneous treatment effect associated with varying baseline risk for the relative scale, a heterogeneous treatment effect for the absolute scale must exist.

For a CBPTI plot, absolute scale is preferred because it provides useful information for clinical settings. An absolute scale gives the actual risk for an individual receiving experimental or standard treatment, but a relative scale gives no information on individual risk. For example, a relative risk reduction of 29.2% corresponds to an absolute risk reduction of 5% vs. 3.5%, and 50% vs. 35.4%. These two scenarios may have different clinical implications if a risk below 5% is considered low and a risk above 5% high.

3.1.2 Statistical Uncertainty

Statistical uncertainty about the treatment effect across subgroups can almost never be ignored. Such information reflects imprecise knowledge about true treatment outcomes and implies the possibility of making an inappropriate decision about which treatment is expected to benefit for a subgroup of patients. A confidence interval/band is often used as part of the graphical presentation to quantify the uncertainty of treatment effects. There are two types commonly displayed in a CBPTI plot, pointwise confidence intervals and simultaneous confidence bands. For a CBPTI plot, Cai et al point out the choice of confidence intervals or bands depends on the clinical purpose [14]. For example, if an author aims to identify a region in which a biomarker is above or below a certain threshold value, pointwise confidence intervals are suggested. If the aim of the study is to evaluate the heterogeneity of the treatment effect, it is important to provide information on the uncertainty of the entire function describing the biomarker treatment interaction. For this purpose, simultaneous confidence bands are recommended.

3.1.3 Benchmarks

The issue of benchmarking is of particular interest. Take ▶Figure 1 as an example. Benchmarking in a CBPTI plot involves

the presentation of a criterion to decide which treatment is better for a specific subgroup. In general, this benchmark is defined by a value which implies no treatment-biomarker interaction, i.e the value at which the difference between two treatment effects is equal to 0 ($\Delta = 0$, $\log(HR) = 0$, $\log(OR) = 0$, $\log(RR) = 0$). Cuzick [2] suggests instead that the value of the overall treatment effect for the trial population should be used as the benchmark. The first option (no difference benchmark) is in the light of counterfactual thinking: which individuals would be better off with the experimental treatment instead of the standard treatment. The second option (mean effect benchmark) stresses the point that the presence of heterogeneity of treatment effect between subgroups is irrelevant to the comparison between experimental treatment and standard treatment within particular subgroups. For a CBPTI plot, benchmarking at mean effect answers the key question of assessing heterogeneity between subgroups.

3.1.4 Clinical Questions

As proposed by Janes et al, a CBPTI plot can aid clinicians in a variety of ways [17]: in choosing one treatment over others for patients on the basis of a biomarker; in indicating what proportion of a population would have a positive response resulting from the treatment selection strategy; or what proportion of patients would have treatment changes after biomarker measurement; or which biomarker is best if several candidates exist. We believe there are still more clinical questions needing to be answered. A CBPTI plot should guide clinicians, their patients, and health policy makers in making good decisions in practice.

3.1.5 Criteria for Assessing a CBPTI Plot

The above aspects are particularly focusing on assessing CBPTI plots. When presenting graphics in clinical research, authors should follow the principles for good plots as suggested by experts [9–11]. Here, we summarize into five principles for future practice when presenting a CBPTI plot.

1. Carefully take into consideration the general principles of visual display.
2. Use absolute units for presenting outcome measures.
3. Display appropriate measures of statistical uncertainty, either a pointwise confidence interval or simultaneous confidence band.
4. For detecting heterogeneous treatment effect across a biomarker's value or making comparisons between biomarkers, a benchmark line should be added for improving direct interpretation.
5. Be intrinsically informative regarding clinical questions relevant to medical decision-making.

3.2 Literature Review

We performed two different literature reviews: first on the development of CBPTI methodology; and second on how CBPTI plots are currently used in clinical research. As CBPTI is rarely mentioned in keywords, titles, or abstracts of papers, we were unable to conduct a systematic review as proposed by the Cochran collaboration. Since the plots of interest do not have a specific name (as opposed to funnel plot, forest plot, and ROC curve), the search could not be done using specific MeSH terms or keywords. Potential papers of interest were not limited to specific diseases or study designs. The formulation of inclusion and exclusion criteria was not feasible. Therefore, the standard searching strategy typically used to limit the retrieval of irrelevant studies from PubMed/MEDLINE is not feasible.

The review of CBPTI methodology was conducted using the strategy of *theoretical saturation*. This is the appropriate analysis "when all categories are well developed in terms of properties, dimensions and variations" [19]. The search stops if "further data gathering and analysis will add little new to the conceptualization, though variations can be discovered" [19]. The use of *theoretical saturation* to guide data collection is based on theoretical sampling. The initial stage of data collection began by identifying some crucial journals relevant to CBPTI. The review extended back to papers published beginning in the year 2000, as the first paper (that we are aware of) relevant to this topic was Gadbury et al in 2000 [20]. This gave a foundation to our search. Since our team has deep insight into this topic, a new idea could be conceptualized and formulated as it emerged from the references being collected. Theoretical sampling continued to be used until no new or relevant references emerged in relation to the categories that had been found. If a reference elicited ideas that had not been heard or that contradicted previous understanding, we extended the literature review to understand this new methodology more fully. The relationship between the categories was well-articulated and each category was densely developed or established before theoretical sampling stopped. Generally, there is no formal requirement to present methods of a theoretical saturation search explicitly since the emphasis is on the conceptual contribution of included literature instead of covering complete and comprehensive articles. Further references that we could uncover would add little new to the original concept. We believe that our report provides a representative result on a basis of the point of theoretical saturation.

For the review of methodology, we selected the top six major biostatistics journals according to Journal Citation Index report in 2014: Statistical Methods in Medical Research (SCI = 4.472), Biostatistics (SCI = 3.072), Methods of Information in Medicine (SCI = 2.248), BMC Medical Research Methodology (SCI = 2.270), Statistics in Medicine (SCI = 1.825), Biometrics (SCI = 1.568). We also selected the first two top clinical trial methodology journals, Trials (SCI = 1.731) and Clinical Trials (SCI = 1.340), for additional methodology review in our report. To avoid missing new proposed methodologies, the survey was extended to search for publications being cited by existing reports and developed by the research groups previously publishing CBPTI methods.

For the second part we limited our search to six major medical journals: New England Journal of Medicine (SCI = 55.873), Lancet (SCI = 42.217), JAMA (SCI = 35.289), Lancet Oncology (SCI = 24.861), Journal of Clinical Oncology (SCI = 18.443), and Annals of Internal Medicine (SCI = 17.810). They are the top

six major medical journals active in the field of randomized controlled trials according to the Journal Citation Index in a 2014 report. A previous report indicated that articles concerning treatment-selection biomarkers are far more likely to appear in high impact journals (24.7% of all RCT articles) than low impact journals (11.6%) [21]. A large proportion of existing CBPTI graphics would thus be found in these high impact journals. A review of the years 2013–2014 for the medical journals was considered sufficient to display representative findings of how researchers use CBPTI plots.

Our survey of methodological and clinical papers was limited to parallel group randomized controlled two armed trials in which an interaction between treatment and a continuous biomarker was discussed. In the survey, for each article we firstly checked the study design. If the design was appropriate, we then restricted attention to CBPTI plots and methods of how they were estimated. Publications providing information on CBPTI without a graphical display were not of interest in our study.

One reviewer (YMS) trained in clinical research methodology extracted the CBPTI plots. If the purpose of a plot was unclear, the doubt was resolved by another senior statistician (UM). All included CBPTI plots were discussed and confirmed by both reviewers (YMS, UM). We critically appraised every CBPTI plot included in our survey regarding the appropriateness of its presentation and clinical utility.

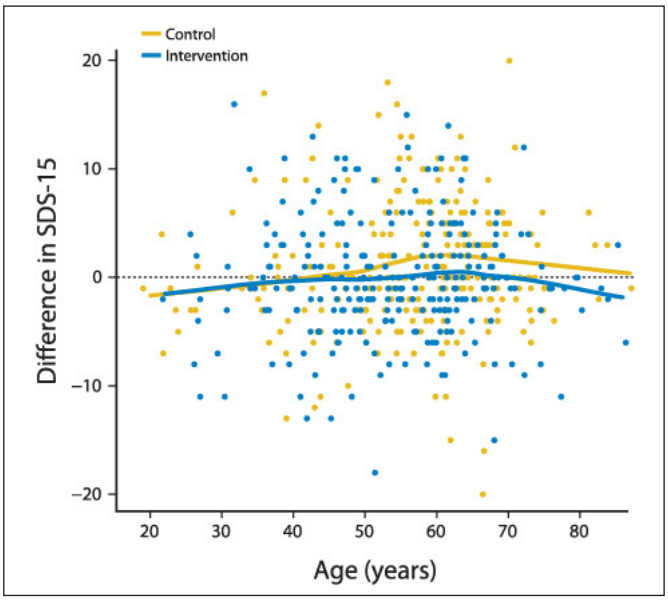

Figure 2
Effect of age on Symptom Distress Scale–15 (SDS-15) score change between baseline and end of study. (Berry et al. JCO. Jan 20, 2014, p. 204) [23]

4. Results

In the biostatistics and clinical trial methodology journals, there were five reports relevant to CBPTI plots during 2000–2014, including two in *Statistics in Medicine* (treatment-effect plot [6] and subpopulation treatment-effect pattern plots [8]), one in *Biostatistics* (Cai et al approach [14]), two in *Biometrics* (Huang's ROC approach [15] and selection impact curve [16]). In addition to these five reports, two newly developed CBPTI plots were found through cross-referencing: one in *Annals of Internal Medicine* (marker-by-treatment predictiveness curves [17]) and one in *International Journal of Biostatistics* (risk curves [22]).

In the medical journals, a total of 767 parallel group RCTs were reported from January 2013 to December 2014: 179 in the *New England Journal of Medicine*, 108 in *The Lancet*, 108 in *The Journal of the American Medical Association*, 26 in *Annuals of Internal Medicine*, 122 in *The Lancet Oncology*, and 224 in the *Journal of Clinical Oncology*. We found five papers [23–27] covering four types of CBPTI plots (two papers presented STEPP. One was selected for our study). Examples of these four types of CBTPI plots will be discussed [23–26]. One plot presented the clinical outcome on a continuous scale, for the remaining three plots the outcome was event data (survival).

We categorized the graphical presentations of these 11 reports into four types of CBPTI plot:

4.1 Distinguishing the Outcome Effect for Each Treatment Group

The classical approach of looking for evidence of an interaction visually is through presenting a so-called *interaction plot*. This type of plot displays the different treatment effects among the groups using separate curves, with treatment effect on the y-axis and biomarker value on the x-axis. The relationship between outcome measure and biomarker can be modelled as a linear or nonlinear function. ▶Figure 2 [23] and ▶Figure 3 [24] are examples of interaction plots found in medical journals.

The aim of ▶Figure 2 is to explore the influence of age on the treatment effect, which is measured as the change in Symptom Distress Scale (SDS)-15 score during the study [23]. The methodology behind the plot was based on nonparametric smoothing techniques. For details see Berry et al. [23]. ▶Figure 2 illustrates the basic aesthetic principles of CBPTI plots. The contrast colors were used to visually differentiate the control group (yellow) and intervention group (blue). The axes are clearly labelled and properly scaled. A legend within the plot helps readers distinguish the colors of treatment groups. The absolute score is used to present the outcome measure. The plot does not show evidence for a striking relationship between age and SDS-15 score change under treatment. It is visible that for age above 50 years the curve corresponding to the intervention is about 2 points in SDS-15 score change below the curve corresponding to the control. Both curves are nearly identical for age below 50. Two crossing predictive lines without interval estimates do not tell readers if there is a significant difference between two groups since the confidence intervals may overlap due to smaller sample size. Often, there is no need to

draw the benchmark in interaction plots. Whether there is an interaction between biomarker and treatment depends on how the two curves deviate from each other. Although the authors present individual outcomes by dots, the distributions are identical. The individual dots provide information on the population variability of treatment outcome; they may be helpful in drawing individual- or population-based conclusions.

►Figure 3 is another example of an interaction plot but extends to evaluate treatment-biomarker interaction under multi-subsets [24]. The authors study potential treatment heterogeneity (fluorouracil (FU) versus FU + oxiliplatin) with respect to the CCRS (colon cancer recurrence score) stratified with respect to cancer stage. The relationship between 5-year risk of recurrence and CCRS is modelled as a linear function. The transformation from the log-scale to the original scale created non-linear relationships. The authors present a formal statistical analysis which is not indicative for interaction, between treatment, stages, and biomarker. The figure can be used to make this complex analysis easily visible. Each curve is estimated by adding interaction terms to the Cox proportional hazards model. On the basis of the principles of visual display, ►Figure 3 shows several limitations. The readers have to read the legend to check which thickness of line belongs to which treatment group. It would be possible to add labels directly to the six lines in a blank area. A total of six lines in a plot make it difficult for the eye to spot any potentially deviant subgroups. To be less complex, the authors could have presented the three stage specific interaction terms of the Cox regression. This way it may have been more obvious if treatment effect heterogeneity is present in each of the three stages. There is essentially no treatment heterogeneity within each stage with respect to the CCRS since the treatment curve and the control curve are parallel for each stage. The plot also presents the distribution of the recurrence score for the entire population as a rug plot alongside the horizontal axis and an estimated normal distribution of scores below the interaction plot. It would have been more informative to answer clinical questions if the authors had shown the distribution of the CCRS through three density plots within each stage (assuming a randomization of the treatment). Again there are no confidence intervals to quantify uncertainty. The plot represents global treatment differences within the three tumor stages, but now biomarker treatment interaction per stage and in total.

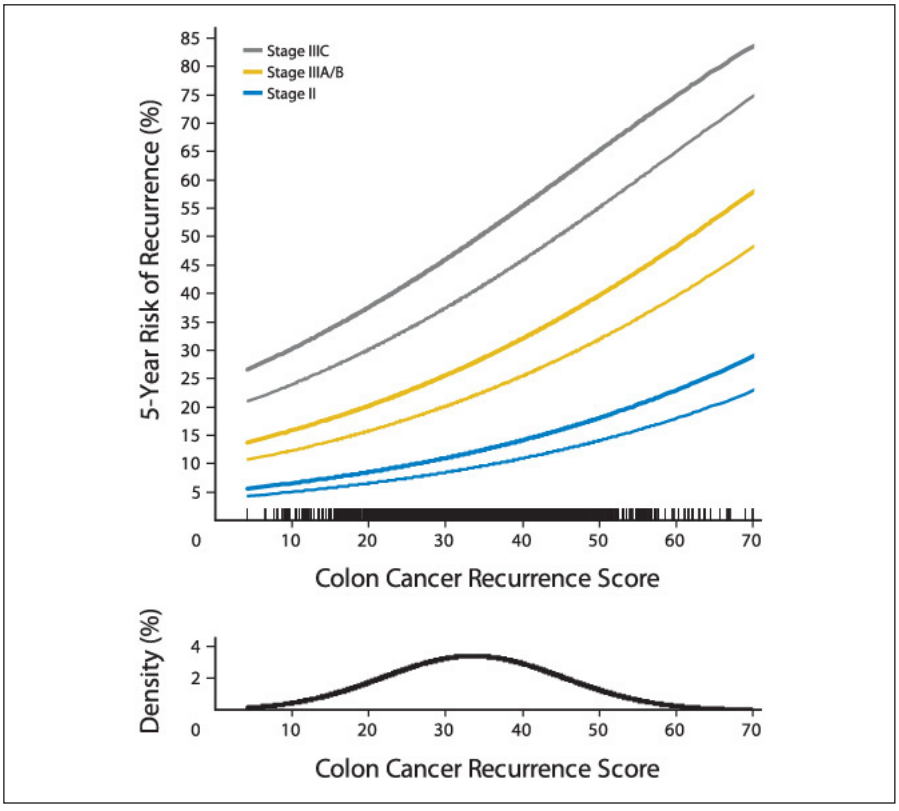

Figure 3 **Relationship between the continuous Recurrence Score (RS) and 5-year recurrence risk by stage and treatment in the National Surgical Adjuvant Breast and Bowel Project C-07.** Thick lines represent fluorouracil (FU)-treated patients, thin lines represent FU + oxaliplatin-treated patients. Blue, gold, and gray colors represent stages II, IIIA/B, and IIIC, respectively. A rug plot depicting the distribution of RS is included at the bottom of the figure, and an estimated normal distribution of scores is provided below. The proportional hazards assumption held (P = .20 for the test of nonzero slope of Schoenfeld residuals *v* time). The relationship between continuous RS and the log hazard of recurrence was linear (P = .84 for the test of nonlinearity). (Yothers et al. JCO. Dec 20, 2013, p. 4515) [24]

4.2 Showing Outcome Difference between Treatment Groups

4.2.1 Partitioning All Individuals into Subpopulations

Bonetti et al were the first to propose the idea of partitioning all individuals into subpopulations on the basis of biomarker value and estimating the treatment differences in each subpopulation [8]. If the line connecting all estimates is not horizontal, there is a heterogeneous treatment effect across the range of the biomarker value. The methodology was briefly explained in paragraph 3 of the background section. In ►Figure 1 for graphical display, there are the lack of the names and scales of axes. The tick intervals should be properly labelled. With absolute outcome measures on y axis, the authors can help readers to indicate what a value above zero or below zero (a ratio above one or below one if presenting hazard ratio on y axis) means by labeling the regions as "Favors tamoxifen plus chemotherapy" and "Favors tamoxifen alone". In the plot, two benchmark lines at the points of no effect and overall effect are displayed. However, it is not clear for readers which benchmark line is used to detect heterogeneity of treatment effect,

even though the authors demonstrated in the original paper that there is no advantage in 5-years DFS for lower values of ER expression when treatment is cyclophosphamide, methotrexate, and 5-fluorouracil (CMF) plus tamoxifen versus tamoxifen alone. The purpose of a benchmark line at overall treatment effect is to allow visual verification of a subgroup's confidence band differing significantly from the overall treatment effect. Further weaknesses of the plot include the outcome unit failing to scale by absolute unit, and that a HR gives no information on individual risk. For therapeutic decision-making, it would be useful to draw a vertical line as a threshold. The threshold could be at the point where the bold line reaches treatment effect for the complete sample. That tells clinicians CMF + tamoxifen is recommended for the patients with the certain range of ER expression. Subpopulation sizes for values of the biomarker are added alongside the x-axis. ▶Figure 4 is a clinical application of STEPP from a medical journal [25] and has similar limitations to those of ▶Figure 1.

Cai et al proposed a more advanced and formally precise approach to presenting outcome difference between treatment groups [14]. They created a score index to group individuals by incorporating subject baseline characteristics and then estimating the treatment difference on a potential outcome framework. Given each score, a spline-based average treatment difference is estimated using a local fitting approach. ▶Figure 5 displays a shaded region and dashed curves to identify two types of uncertainty estimates [14]. The aim of the plot is to detect if a patient's change in CD4 count from the baseline level to week 24 differs across the individual's score index when comparing a 3-drug combination with a 2-drug combination. The authors demonstrate that the change in CD4 count from baseline to week 24 is consistent at lower scores but increases significantly for scores above 50. A horizontal benchmark line at the average treatment difference would be recommended for the plot to improve direct interpretation. The plot could not answer clinical questions about which treatment would influence a therapeutic decision, since the 3-drug combination always performs better than the 2-drug combination over the scores. Is the higher adverse event risk under the new treatment compensated by better response to treatment?

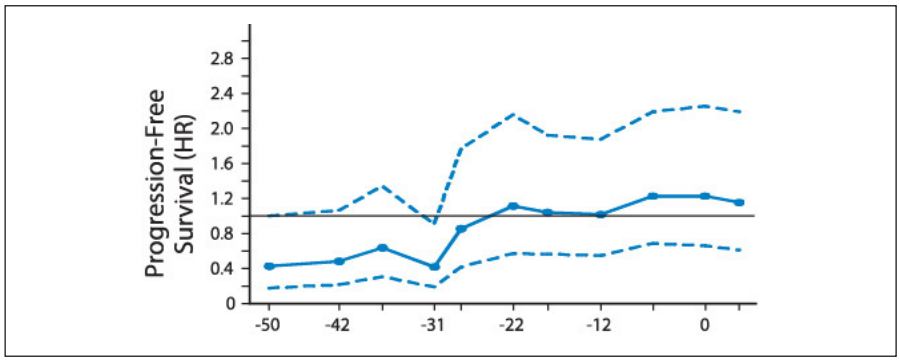

Figure 4 Sliding-window subpopulation treatment effect pattern plot analysis of the treatment effect of adding cetuximab to chemotherapy in patients with KRAS wild-type as measured by hazard ratios (HRs) for progression-free survival (chemotherapy plus cetuximab v chemotherapy alone. HR values < 1 suggest benefit of adding cetuximab, with 95% CIs in dashed lines. The x-axes indicate median tumor shrinkage at 8 weeks for patients in each of the overlapping subpopulations. (Piessevaux et al. JCO. Oct 20, 2013, p. 3770) [25]

4.2.2 Modelling the Functional Form of the Interaction

With the advance of the use of regression models, methodological studies relevant to CBPTI plots are focused on modelling the functional form of the interaction between a continuous biomarker and treatment in a multivariable regression setting. Take the simple example from ▶Figure 2. One could portray a *contrast plot* which simply presents the interaction component of the linear regression model. In terms of the functional form of the interaction between a continuous biomarker and treatment, one fits $\Delta(\text{score}) = \alpha + \beta \cdot \text{age} + \gamma \cdot \text{treat} + \delta \cdot \text{treat} \cdot \text{age}$, and presents the contrast line as $f(\text{age}) = \gamma + \delta \cdot \text{age}$. For general cases, the straight line function is simple and may be adequate. However, it may lead to loss of power and give faulty interpretation if a non-linear relationship is incorrectly assumed to be linear [6]. Normally, a contrast plot cannot display individual data because it presents a mean difference between the treatment groups. The parallel-group design does not provide the outcome of both treatments for the same individual: Since an individual patient only belongs to one group, there is no natural counterpart for

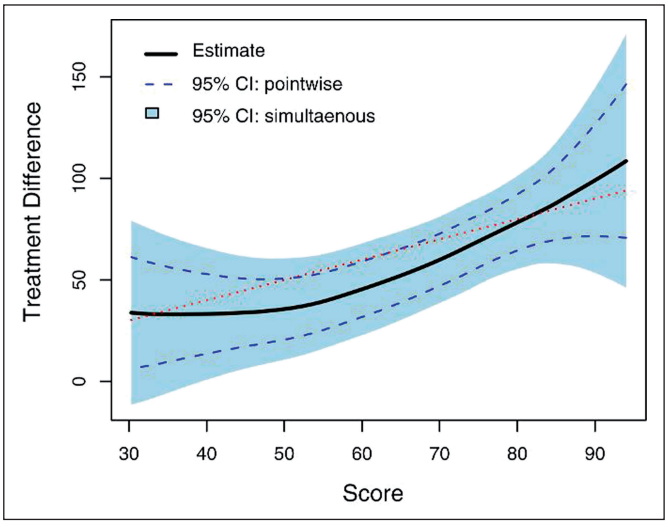

Figure 5 Estimated treatment differences (thick curve), 3-drug combo minus 2-drug combo, with respect to week 24 CD4 changes over the score and the corresponding 95% pointwise (dashed curve) and simultaneous (shaded region) confidence intervals. (Cai et al. Biostatistics. Apr 2011, p. 277) [14]

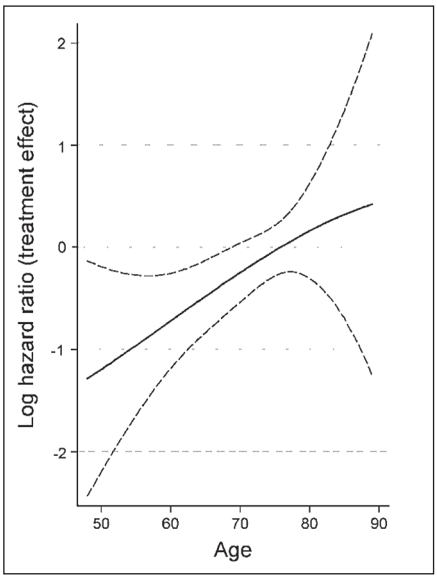

Figure 6 Prostate cancer data: treatment × age interaction: the effect of treatment by age, with 95% pointwise confidence interval. Functions were estimated in multivariable adjustment models and fitted using FP2 functions with powers (3; 3). (Royston et al. Stat Med. Aug 30, 2004, p. 2516) [6]

calculating a difference. The difference could only be given if for single patients their counterfactual outcomes under the second treatment could be known. Recent approaches have tried to estimate individual treatment effects within a counterfactual framework [15, 28]. This could be presented in a corresponding contrast plot.

There are a variety of approaches to modelling the interaction between a continuous biomarker and treatment as a non-linear function [6, 29, 30]. Royston and Sauerbrei proposed the use of a power transformation, termed "fractional polynomial", in modeling the functional form of a continuous biomarker [6]. The powers $S = \{-2,-1,-0.5,0,0.5,1,2,3\}$ are suggested. The functional form of continuous biomarker can be formalized by either first-degree or second-degree fractional polynomial function. For each treatment group, interaction between a continuous biomarker and treatment is modelled by a fractional polynomial with the same powers but different regression coefficients. Then the difference between two functions for each treatment group is calculated and tests for significance. ▶Figure 6 is an example of a CBPTI plot based on their approach, named "treatment-effect plot" [6]. The outcome measure is given as the log hazard ratio, which is not easily understandable in clinical settings and hazard ratio provides no information on the likelihood that an individual would benefit. In the original paper, the authors revealed that treatment E is favorable for the 50–75 age group, but may be harmful for ages over 80. It would be suggested to show a benchmark line at overall treatment effect level for correct and fast interpretation.

Harrell proposed an approach of employing restricted cubic splines in modelling the functional form of a continuous covariate [30] which is also helpful to model the interaction between a continuous biomarker and treatment: A functional form of continuous biomarker is fitted by restricted cube spline with knot $k = \{0,3,4,5,6\}$ and includes them as main effect terms and as treatment interaction terms. Akaike's information criterion is used for the selection of k. Having determined the spline function for the interaction, it is possible to calulate the corresponding simultaneous confidence intervals. Other proposed existing spline methods could be employed in modelling interaction between a continuous biomarker and treatment as well. ▶Figure 7 is a clinical example from a medical review [26]. The approach behind the plot is the relation between imputed anti-circumsporozoite antibodies and protection against malaria in a Cox proportional hazards model with cubic spline function. The "upside down" J-curve beyond the no-effect point clearly shows that there is heterogeneity of protection effect across the range of anti-circumsporozoite antibodies titer. However, the benchmark line at the no-effect point (i.e., a hazard ratio of 1) leads to false interpretation. The interpretation of the heterogeneous protection effect would be correct if this benchmark line were moved to a horizontal line at the overall treatment effect level. Although the authors indicate that there is reduced risk of clinical malaria with increasing antibody titers at values above 1000 enzyme-linked immunosorbent assay unit (EU) per milliliter in the legend, a visual display of the threshold of protection change would be very useful for clinicians. However, one may be more optimistic regarding the decision for the placement of the threshold by positioning it where the fitted regression curve reaches the no-effect point (around 105 EU/ml).

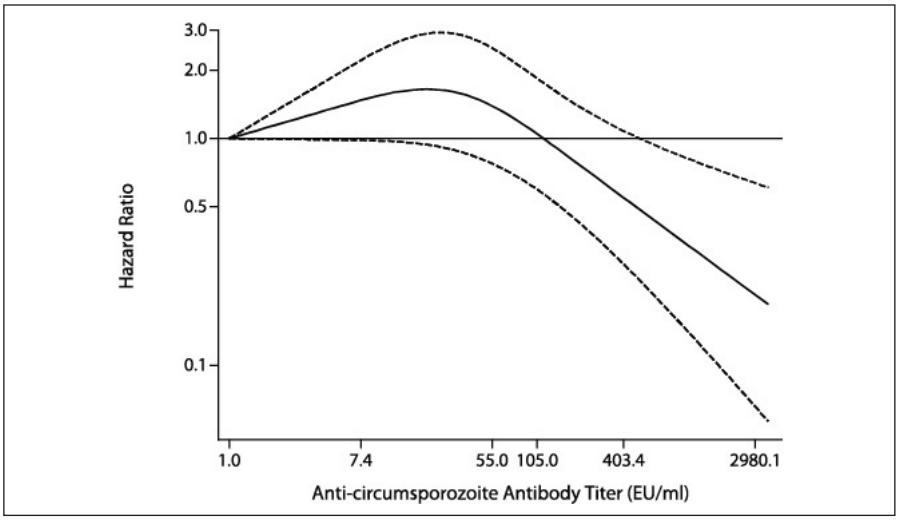

Figure 7 The association between imputed anti-circumsporozoite antibody titers and the hazard ratio for clinical malaria episodes among children who received the RTS,S/AS01E vaccine, according to a Cox regression model with cubic splines and with a baseline titer of 1.0 enzyme-linked immunosorbent assay unit (EU) per milliliter as the reference. The dotted lines indicate the 95% confidence interval. There was no significant variation in risk between 1 EU per milliliter and 1000 EU per milliliter (i.e., the confidence intervals include a hazard ratio of 1.0); at values above 1000 EU per milliliter, however, there was a reduced risk of clinical malaria with increasing antibody titers. (Olotu et al. NEJM. March 21, 2013, p. 1119) [26]

In a majority of studies, the assumption of a linear function of continuous biomarker may be satisfied. However, in some cases continuous biomarker may represent a non-linear relationship with outcome. For clinical practice, the use of a spline function is helpful to explore the heterogeneity of treatment effect; however, fractional polynomials should be used in the final model [31, 32].

Hazard ratios are presented in ▶Figures 1, 4, 6 and 7, and may be no good way to interpret quantity. We can transform a hazard ratio into risk probability [33].

$$\Pr(T_{experimental} < T_{standard}) = \frac{\exp\{\gamma + \delta \cdot biomarker\}}{1 + \exp\{\gamma + \delta \cdot biomarker\}}$$

where T denotes survival time, and γ and δ are the regression coefficients for treatment effect and interaction term, respectively. The risk probability can be interpreted in a counterfactual way as a patient under the experimental treatment experiencing the event before a patient under the control treatment. The idea is also corresponding to concordance probability [34], defined as the risk of event that a pair is concordant if one with experimental treatment has the first event. An example is presented in ▶Figure 8. The plot shows that the risk probability of unfavorable treatment effect due to interferon-alpha treatment increases with increasing white blood cell count. Detailed information on the dataset was documented Royston et al in 2004 [35]. ▶Figure 8 presents many good practices: the outcome is scaled as absolute units, measures of uncertainty are shown, and there is a horizontal line as a benchmark for the decision of the threshold. A threshold where the bold curve reaches the point of 50% of the population is recommended. ▶Figure 8 could answer clinical questions as how probable it is to experience under experimental treatment the event before standard treatment.

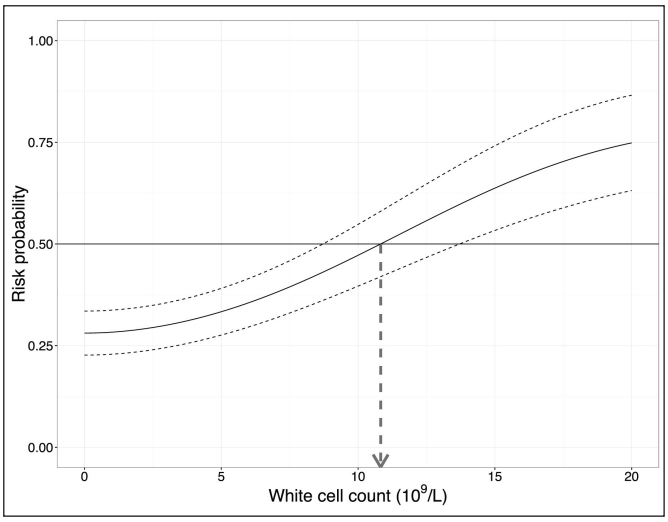

Figure 8
Risk probability for a patient with interferon-alpha treatment experiencing the event before a patient with medroxyprogesterone acetate treatment. (The dataset was obtained from a Medical Research Council RE01 phase III randomized controlled trial [35])

4.3 Evaluating the Proportion of Population Impact of the Biomarker

The major limitation of interaction plots and contrast plots is its lack of information for medical decision-making since the predictive capacity of a biomarker should be concerned with the population impact of treatment selection [16–17]. The methodologies relevant to the proportion of population impact of the biomarker include the selection impact curve [16], the marker-by-treatment predictiveness curve [17], and the risk curve [22].

The selection impact curve was proposed to display the response rate (the proportion of population who benefit from experimental treatment if their biomarker value exceed cutoff but to assign them to standard treatment otherwise) as a function of treatment assignment based on the biomarker, as shown in ▶Figure 9 [16]. For health-decision makers, it is helpful to identify the proportion of population that are more likely to benefit from assigned treatments. On the other hand, the selection impact curve has similar property as ROC curve since both axes are scaled as percentile and there is a tradeoff between them. The plot allows making comparison between candidate biomarkers. The best biomarker for treatment-selection is the concave downward curve that is the closest to the point of (1, max(response rate)). There are several improvements could be made to these plots to increase their clinical suitability. The axes are poorly labelled for clinical settings and difficult to understand for most clinicians, a result of the paper being published in a biostatistics journal, the readers of which being primarily statisticians. Confidence bands displaying statistical uncer-

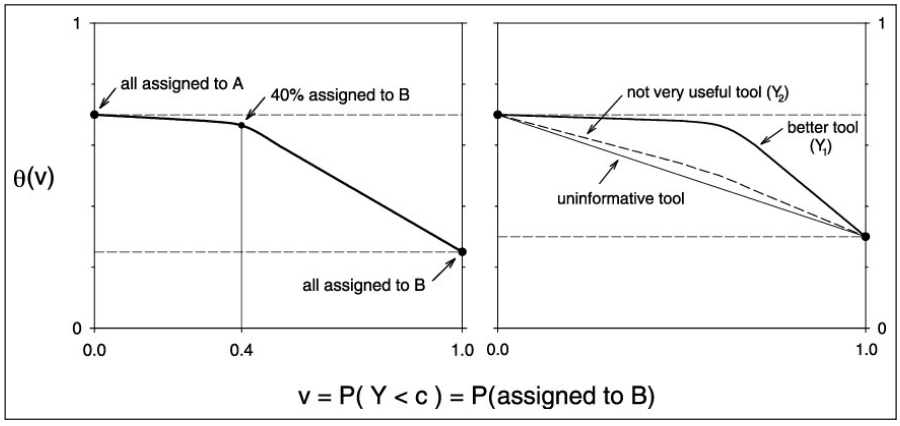

Figure 9 **A schematic diagram of the selection impact (SI) curve, ϑ(v) = the population response rate = $P\{D = 1 | (Y > c, T = 1)$ or $(Y < c, T > 0)\}$.** (Song et al. Biometrics. Dec, 2004, p. 875) [16]

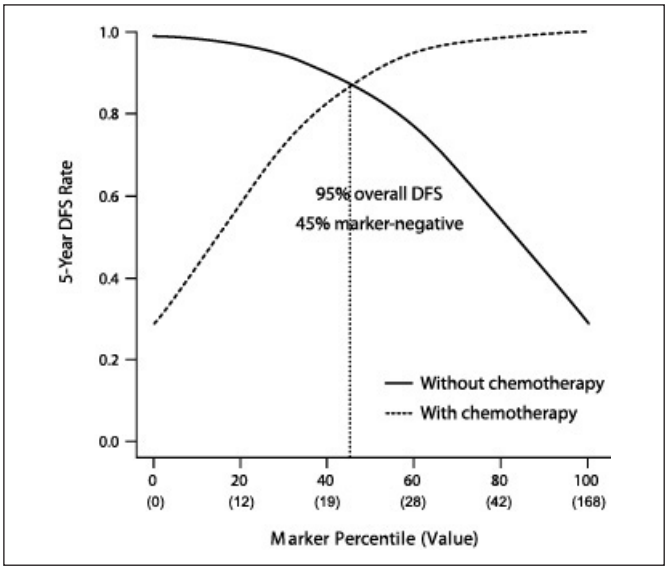

Figure 10 **The 5-year disease-free survival (DFS) rate plotted as a function of marker percentile, with raw marker values shown in parentheses.** The overall DFS rate with use of the marker for guiding treatment is shown, as well as the percentage of women who have higher DFS rates with tamoxifen alone (marker-negative). (Janes et al. Ann Intern Med. Feb 15, 2011, p. 255) [17]

tainty are encouraged to ease comparability.

The other innovative CBPTI plots proposed by the same research group, the marker-by-treatment predictiveness curve (►Figure 10) [17] and the risk curve, (►Figure 11) [22], have similar properties. The principle is to illustrate the expected treatment benefit or probability of a certain outcome given a specific biomarker value that is presented in a corresponding interaction plot. The advantage of both graphic presentations is the x-axes are also scaled as percentiles and additional information can be derived if the population distribution of the biomarker is known. Both plots are clearly labelled and the outcome measures are scaled as the same absolute units, allowing easy comparison between candidate biomarkers. The performance of the biomarker depends on how the two curves deviate from each other: larger variation implies a better performance of the biomarker in differentiating the treatment effect. For multiple comparisons across biomarkers, however, it is not recommended to display the curves in the same plot since this can inhibit ease of readability. A good approach is to present the difference between two treatments for each candidate biomarker by using different colors in one plot (as presented in figure 2 in Janes et al paper [22]). All biomarkers share the same benchmark line at overall treatment effect level. Again, the confidence intervals for statistical uncertainty are missing in ►Figure 10.

Here we proposed the *proportion of unfavorable treatment effect plot* that estimates the proportion of subjects in a population who have unfavorable outcome due to experimental treatment. The approach uses parametric simulation. Given each biomarker value, the arithmetic mean of predicted difference between groups is estimated and then computes the probability that a normally distributed random sample will be above than that zero (the event risk of event of experimental treatment minus that of standard treatment). The example is presented in the ►Online Appendix.

4.4 Showing Classification Accuracy of Biomarker

The ROC curve can be used as a graphical method of distinguishing poor or good responders to a new treatment. This approach is motivated by the fact that the former CBPTI plots are highly dependent on the scales which the continuous biomarkers are set. When comparing multiple biomarkers with difference scales, there is insufficient evidence to assess the performance of biomarkers. Huang et al proposed a new approach, ROC curve, which puts candidate biomarkers on the same scale to facilitate comparisons [15]. Their ROC curve is constructed under strict assumptions on the basis of a potential outcome framework. We propose an approach for normal distributed endpoints which can be applied to parallel group RCTs. Although there are some similarities to the selection impact curve, a ROC curve provides the sensitivity and 1-specificity for the performance of biomarker in distinguishing good or poor responders.

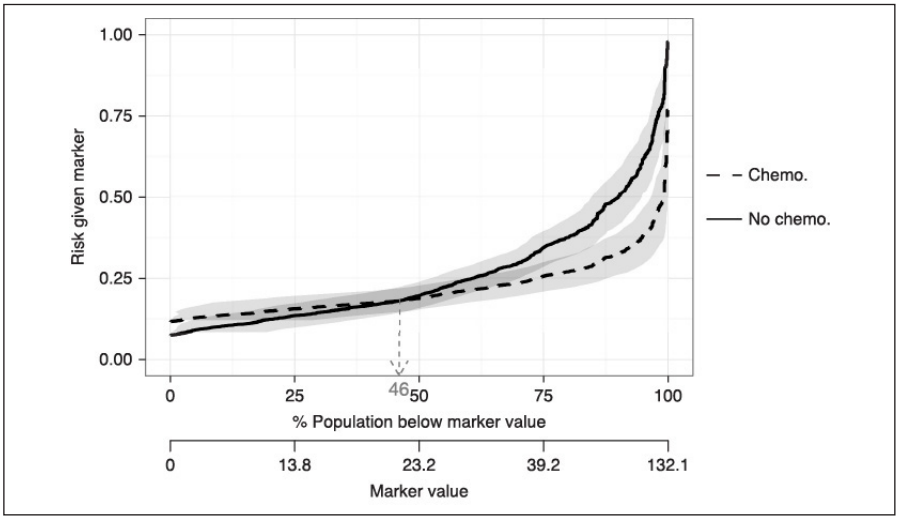

Figure 11 **Risk of 5-year breast cancer recurrence or death as a function of treatment assignment and marker percentile, for the Oncotype-DX-like marker.** Horizontal pointwise 95% confidence intervals (CIs) are shown. Forty-six percent of women have negative treatment effects according to the Oncotype-DX-like marker; these women can avoid adjuvant chemotherapy. (Janes et al. Int J Biostat. 2014, p. 102) [22]

Given a particular cut-off value z_0 of the biomarker Z, an individual can be classified as a good responder under a new treatment ($\Delta < 0$) if Z is above the threshold z_0. Here, Δ may be the log(HR) expressing the event risk of the new treatment relative to the standard treatment for the chosen individual. The classification accuracy of the treatment-selection biomarker is characterized in two ways: the true positive fraction (TPF), defined as the probability of correctly identifying a good responder, and the false positive fraction, (FPF) defined as the probability of incorrectly classifying a bad responder as a good responder. Formally, $TPF(z_0) = P[Z > z_0 \,|\, \Delta < 0]$ and $FPF(z_0) = P[Z > z_0 \,|\, \Delta \geq 0]$. A ROC curve for varying cutoff values is drawn with the TPF on the vertical axis and the FPF on the horizontal axis: $ROC(x) = TPF(FPF^{-1}(x))$, $0 < x < 1$. ▶Figure 12 comes from Huang's approach but similar to our idea for two candidate treatment-selection biomarkers [15]. Ad5 is a better biomarker than age for treatment selection because it deviates more from the diagonal. This plot is intrinsically informative for comparing candidate biomarkers in a clinical setting. Methodological limitations include lack of confidence bands elucidating uncertainty estimates, and the lack of the diagonal line which can serve as a benchmark.

The ROC curve describes how well the good responders can be differentiated from the bad responders based on some biomarker measurement (high marker values indicating good response). However, sensitivity and specificity are of no practical use when it comes to helping clinicians estimate the probability of good response in individual patients since both quantities elucidate the distributions of the biomarker in responder groups and do not directly provide information on individual prediction. Providing positive predictive values (PPV) and negative predictive values (NPV) is of highest interest: $PPV(z_0) = P[\Delta < 0 \,|\, Z > z_0]$ and $NPV(z_0) = P[\Delta \geq 0 \,|\, Z \leq z_0]$. The calculation of both values is a function of the constitution of the entire sample. The interpretation of PPV and NPV are the proportions of good responders and bad responders in biomarker measurement that are above and below the cut-off value, respectively. However, any approach that uses the conditional probability of outcome given biomarker cutoff is defective for individual prediction. Therefore, to understand the implications, on the basis of our ROC approach, we calculate the proportions of good responders (GR) and bad responders (BR) given the biomarker value. The *prediction curve* is helpful for predicting the probabilities of good response and bad response given an individual's biomarker value. The plot is displayed in the ▶Online Appendix.

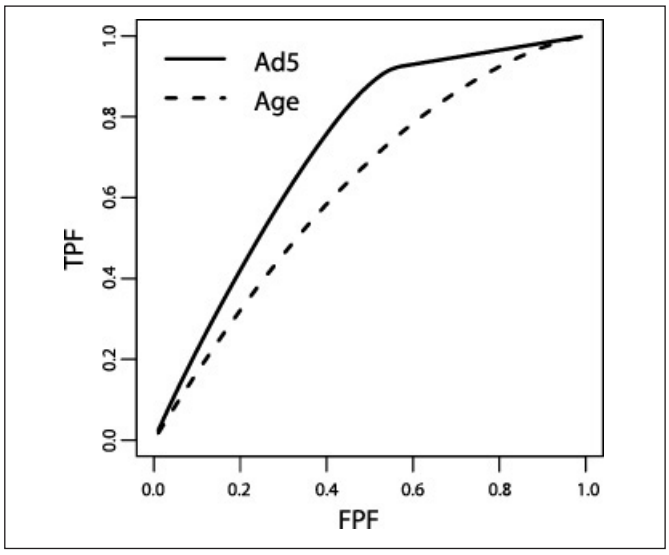

Figure 12 **Plot of ROC curves for classifying a subject into treatment-effective or treatment-ineffective groups.** (Huang et al. Biometrics. Sept, 2012, p. 694) [15]

5. Discussion

We propose a set of criteria which help to create clear and informative CBPTI plots, such as general principles of visual display, appropriate quantification of statistical uncertainty, use of units presenting absolute outcome measures, correct display of a benchmark, and informative content to answer clinical questions. They are in consonance with ideas formulated previously by various authors and are compiled for the first time in the proposed list. ▶Table 1 summarizes our assessment for each CBPTI plot based on our principles. The proposal is open for discussion.

In order to assess the usefulness and completeness of the criteria list, we performed two literature reviews, one oriented toward methodology, the other focused on the present practice documented in medical journals. We found that newly developed methodological approaches to CBPTI plots are an attempt to answer clinical questions relevant to medical decision-

Table 1 The summary checklist for assessing CBPTI plots based on the guiding principles.

	Fig. 1	Fig. 2	Fig. 3	Fig. 4	Fig. 5	Fig. 6	Fig. 7	Fig. 8	Fig. 9	Fig. 10	Fig. 11	Fig. 12
1. Following the principles of visual display	No	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
2. Using absolute unit for outcome measures	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes
3. Quantifying statistical uncertainty	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No
4. Displaying correct benchmark line	No	--	--	No	No	No	No	Yes	Yes	--	--	No
5. Answering clinical questions	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

making but are not commonly employed in clinical research reporting practice. Although it may take time for the translation of the innovative methodologies from biostatistics journals to medical journals, the main reason for their lack of use may be the complex algorithms and the unavailability of the statistical software necessary for their implementation. Moreover, we found that the utilization of CBPTI plots in medical journals is burdened by poor debatable practice: the outcome measure being given as relative unit instead of absolute unit, the lack of presentation of statistical uncertainty, incorrect benchmarking, and not answering clinical questions. We encourage researchers to follow the principles specified in our study for improved presentation of graphics in future trial reports.

We examined the following four types of CBPTI plots. 1.) Explorative interaction plots which show the biomarker effect in each treatment group on outcome. The biomarker effect may also be explored in specific subgroups. Such plots may also be combined with individual patient data showing dots for each individual outcome. 2.) Contrast plots which illustrate statistical interaction concepts and the analyzed heterogeneity of the biomarker on the treatment effect. In general these plots do not communicate estimates of probabilities for the individual patient's advantage under a specific treatment choice. But in principle this could be easily achieved by a small computational effort. 3.) Plots showing the population impact of a biomarker on the treatment effect. These plots are conceptually linked to the contrast plots. 4.) ROC based plots which show the potential of a biomarker to discriminate between good and bad responders. They are very convenient to compare the discriminatory potential of biomarkers regarding patient-treatment interaction. These plots can be combined via Bayes' theorem with the specific prevalence of biomarker levels in a population to derive predictive values which are of high interest for clinical decision-making. These relationships are presented in ▶ Figure 13.

In 2014, the European Medicines Agency (EMA) proposed the new guidelines focused on the investigation of subgroups in clinical research [36]. Section 4.3 addressed the basic considerations for evaluation of heterogeneous treatment effect and associated data presentations. However, the principles ignore some key issues, such as confidence intervals and benchmarks, that are relevant to interpretation of results and medical decision-making in data presentation. Their guidelines need to be further discussed and improved.

First, the EMA recommends showing statistical uncertainty using confidence intervals. The confidence intervals for binary or categorical biomarkers are straightforward, while confidence bands for continuous biomarker setting need more clarification.

Second, the EMA states that a forest plot is a useful tool in investigation of treatment-covariate interaction. However, one crucial issue relevant to direct interpretation of heterogeneous treatment effect is missing. Drawing a benchmark line for direct interpretation, either at the point of overall treatment level or at the point of no treatment level, is of critical importance. The benchmark line should be at the overall treatment effect since we are not interested in the comparison between experimental treatment and standard treatment but in the heterogeneity of treatment effect among subgroups. Incorrect benchmarking impairs medical decision-making. Equivalent considerations apply also to graphical presentation in treatment-continuous biomarker interaction.

Third, both ICH E9 and the new guidelines on the investigation of subgroups proposed by the EMA indicate heterogeneity of treatment effect should be detected first through the addition of interaction terms to the regression models. The new guidelines further point out that reporting only the interaction term is inadequate. It is recommended to show differences in treatment effects between subgroups. However, the guidelines for investigation of heterogeneous treatment effect are still insufficient. Huang et al [15] give the example of two scenarios giving the same regression coefficient for an interaction term but very different biomarker performance due to scale and functional form of biomarker.

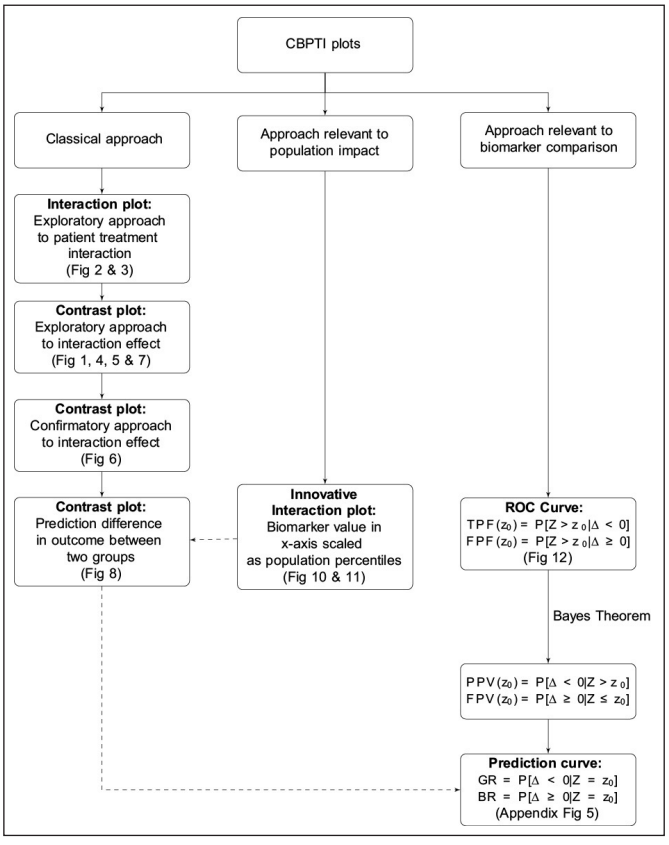

Figure 13
Flow diagram for the development of CBPTI plot.

Interaction plots or contrast plots fail to lead to medical decision-making. Huang [15] proposes a ROC curve to overcome this limitation, as it provides a natural common scale for comparing true positive fraction and false positive fraction achieved with treatment selection policies based on candidate biomarkers. However, Huang's approach is used for binary outcome and is constructed under a potential outcomes framework with severe restrictions. They assume the experimental treatment will either not be harmful or will have not any benefit. The limitation on the use of their approach in general settings can be anticipated. Therefore, we proposed a new ROC curve which is used for continuous outcome or survival time. Under a randomization assumption, we assume biomarker distributions among treatment groups are equal. Thus, treatment differences given biomarker values can be estimated. The approach is straightforward and can be applied in any randomized controlled trial with parallel group design.

There are some limitations and strengths of this study. First, we provide a list of criteria which is based on our personal experience and knowledge of the methodological literature. Further aspects may be added. Second, we performed a formal search strategy following the principles of Cochrane reviews neither in the biostatistical/methodological journals nor in the medical journals. Instead a hand search was performed based on the principle of theoretical saturation. Our survey was an attempt to review how researchers present CBPTI plots, which are used to combine findings in clinical papers. The current search engines fail to conduct a sensible search for our purposes. However, we made great efforts to locate the existing CPBTI plots. We know there are papers which provide CBPTI plots in journals which were not searched. For example, reference [37] provides two plots presenting the functional form of the interaction between biomarker and treatment. We believe that the report can provide a comprehensive and representative result on how CBPTI plots are used in the reporting practice of major medical journals. Third, the search was limited to randomized controlled trials with parallel group design. In principle, the concepts of CBPTI cannot be employed in observational or registry studies. The randomization is crucial in ensuring adequate distributions of the biomarker values in the control and experimental groups. Since clinical practice is not randomized it is an open problem how these plots actually support practical clinical decision-making (a question of internal/external validity). In spite of the limitations, our review provides not only comprehensive methodologies on assessing CBPTI plots but also critical principles of reporting CBPTI for improved future study. This work promotes the development of personalized medicine in the clinical setting.

We have developed a first version of an R package called cbpti, which implements interaction plots, contrast plots, proportion of unfavorable treatment effect plots, ROC curves, and prediction curves. For detailed explanations, please see the ▶Online Appendix. The main limitation of our R package is the restriction to a linear biomarker-treatment effect relationship. The extension to a nonlinear functional form for the biomarker-treatment interaction is under development [38]. It is also necessary to extend the functionality of the package to binary and survival outcomes. This is the plan for future research.

The current package is available in the ▶Online Appendix. Additional graphical methods for risk curves and Huang's ROC curve implemented by R software packages are also available at http://labs.fhcrc.org/janes/index.html and http://labs.fhcrc.org/huang /index.html, respectively.

6. Conclusion

Evaluating the interaction between treatment and a continuous biomarker requires advanced statistical methodology, which makes formal communication of the results for the clinical setting difficult. Graphical presentation may be particularly informative for a researcher who is not an expert in biomarker statistics. Many CBPTI plots are presented in our report. Although interaction plots and contrast plots are commonly used in medical literature, we would encourage researchers to employ new methods such as the proportion of unfavorable treatment effect plot, selection impact curve, modified marker-by-treatment predictiveness curve, modified risk curve, ROC curves, and prediction curves, as such approaches fulfill the principles and answer key clinical questions relevant to medical decision-making. The proposed principles in our report would be helpful for the improved presentation of CBPTI plots in future practice.

Abbreviations

AUC: Area under the Curve; BR: Bad responder; CBPTI: Continuous biomarker in patient treatment interaction; CCRS: Colon cancer recurrence score; CI: Confidence interval; CMF: cyclophosphamide, methotrexate, and 5-fluorouracil; DFS: Disease-free survival; EU: Enzyme-linked immunosorbent assay unit; ER: estrogen receptor; FPF: false positive fraction; FU: Fluorouracil; GR: Good responder; HR: hazard ratio; RCTs: Randomized controlled trials; RS: Recurrence score; ROC: Receiver operating characteristic; SSD: Symptom distress score; STEPP: Subpopulation treatment effect pattern plot; TPF: True positive fraction.

Author Contribution

YMS and UM conceptualized and designed the study. YMS reviewed the literature and drafted the manuscript. YMS and UM evaluated all CBPTI plots. UM gave critical revision of the manuscript for important intellectual content. RW polished manuscript. YMS, LDL and UM performed statistical analysis and developed the R package. YMS, LDL, RW and UM read and approved the final manuscript.

Acknowledgment

▶Figures 2, 3, & 4 are reproduced with permission from the *Journal of Clinical Oncology* and originally published by the *American Journal of Clinical Oncology*. [Figure 2: Berry DL, Hong F, Halpenny B, Partridge AH, Fann JR, Wolpin S, Lober WB, Bush NE, Parvathaneni U, Back AL, Amtmann D, Ford R: Journal of Clinical Oncology Vol. 32 (Issue 3), date: January

20, 2013]. [Figure 3: Yothers G, O'Connell MJ, Lee M, Lopatin M, Clark-Langone KM, Millward C, Paik S, Sharif S, Shak S, Wolmark N: Journal of Clinical Oncology Vol. 31 (Issue 36), date: December 20, 2013]. [Figure 4: Piessevaux H, Buyse M, Schlichting M, Van Cutsem E, Bokemeyer C, Heeger S, Tejpar S: Journal of Clinical Oncology Vol. 31 (Issue 30), date: October 20, 2013].

▶Figures 1 & 6 are reproduced with permission from *Statistics in Medicine* by John Wiley & Sons, Ltd. [Figure 1: Bonetti M, Gelber RD: Statistics in Medicine Vol. 19 (Issue 19), date: October 15, 2000] [Figure 6: Royston P, Sauerbrei W: Statistics in Medicine Vol. 23 (Issue 16), date: August 30, 2004].

▶Figure 5 is reproduced with permission from *Biostatistics* by Oxford University Press. [Figure 5: Cai T, Tian L, Wong PH, Wei LJ: Biostatistics Vol. 12 (Issue 2), date: September 28, 2010].

▶Figure 7 is reproduced with permission from the *New England Journal of Medicine* by Massachusetts Medical Society. [Figure 7: Olotu A, Fegan G, Wambua J, Nyangweso G, Awuondo KO, Leach A, Lievens M, Leboulleux D, Njuguna P, Peshu N, Marsh K, Bejon P: the New England Journal of Medicine Vol. 368 (Issue 12), date: March 21, 2013].

For ▶Figure 8, the dataset was provided by Prof. Willi Sauerbrei.

▶Figures 9 & 12 are reproduced with permission from *Biometrics* by John Wiley & Sons, Inc. [Figure 9: Song X, Pepe MS: Biometrics Vol. 60 (Issue 4), date: December of 2004]. [Figure 12: Huang Y, Gilbert PB, Janes H: Biometrics Vol. 68 (Issue 3), date: September of 2012].

▶Figure 10 is reproduced with permission from the *Annals of Internal Medicine* by the American College of Physicians. [Figure 10: Janes H, Pepe MS, Bossuyt PM, Barlow WE: Annals of Internal Medicine Vol. 154 (Issue 4), date: February 15, 2011].

▶Figure 11 is reproduced with permission from *The International Journal of Biostatistics* by De Gruyter. [Figure 11: Janes H, Brown MD, Huang Y, Pepe MS: The International Journal of Biostatistics Vol. 10 (Issue 1), date: April 2, 2014].

This study was not funded by special institutes.

Conflict of Interest

The authors declare that they have no competing interests.

References

1. Van Cutsem E, Kohne CH, Hitre E, Zaluski J, Chang Chien CR, Makhson A, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. N Engl J Med. 2009; 360: 1408–1417.
2. Cuzick J. Forest plots and the interpretation of subgroups. Lancet. 2005; 365: 1308.
3. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006; 332: 1080.
4. Eisner MD. The challenge of subgroup analyses. N Engl J Med. 2006; 355: 211; author reply 211–2.
5. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994; 86: 829–835.
6. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med. 2004; 23: 2509–2525.
7. Royston P, Sauerbrei W. Interactions between treatment and continuous covariates: a step toward individualizing therapy. J Clin Oncol. 2008; 26: 1397–1399.
8. Bonetti M, Gelber RD. A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. Stat Med. 2000; 19: 2595–2609.
9. Pocock SJ, Travison TG, Wruck LM. Figures in clinical trial reports: current practice & scope for improvement. Trials. 2007; 8: 36.
10. Pocock SJ, Travison TG, Wruck LM. How to interpret figures in reports of clinical trials. BMJ. 2008; 336: 1166–1169.
11. Tufte ER. The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press; 1983.
12. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet. 2005; 365: 256–265.
13. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010; 340: c117.
14. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12: 270–282.
15. Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics. 2012; 68: 687–696.
16. Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. Biometrics. 2004; 60: 874–883.
17. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med. 2011; 154: 253–259.
18. Viale G, Giobbie-Hurder A, Regan MM, Coates AS, Mastropasqua MG, Dell'Orto P, et al. Prognostic and predictive value of centrally reviewed Ki-67 labelling 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. 2008; 26: 5569–5575.
19. Corbin J, Strauss A. Basic of qualitative research: techniques and procedure for developing ground theory. 3rd ed. Thousand Oaks: SAGE Publications; 2008.
20. Gadbury GL, Iyer HK. Unit-treatment interaction and its practical consequences. Biometrics. 2000; 56: 882–885.
21. Sun X. Briel M, Busse JW, You J, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomized controlled trials: systematic review. BMJ. 2012; 344: e1553.
22. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. Int J Biostat. 2014; 10: 99–121.
23. Berry DL, Hong F, Halpenny B, Partridge AH, Fann JR, Wolpin S, et al. Electronic self-report assessment for cancer and self-care support: results of a multicenter randomized trial. J Clin Oncol. 2014; 32: 199–205.
24. Yothers G, O'Connell MJ, Lee M, Lopatin M, Clark-Langone KM, Millward C, et al. Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. J Clin Oncol. 2013; 31: 4512–4519.
25. Piessevaux H, Buyse M, Schlichting M, Van Cutsem E, Bokemeyer C, Heeger S, et al. Use of early tumor shrinkage to predict long-term outcome in metastatic colorectal cancer treated with cetuximab. J Clin Oncol. 2013; 31: 3764–3775.
26. Olotu A, Fegan G, Wambua J, Nyangweso G, Awuondo KO, Leach A, et al. Four-year efficacy of RTS,S/AS01E and its interaction with malaria exposure. N Engl J Med. 2013; 368: 1111–1120.
27. Wolff AC, Lazar AA, Bondarenko I, Garin AM, Brincat S, Chow L, et al. Randomized phase III placebo-controlled trial of letrozole plus oral temsirolimus as first-line endocrine therapy in postmenopausal women with locally advanced or metastatic breast cancer. J Clin Oncol. 2013; 31: 195–202.
28. Laubender RP, Mansmann U. Estimating individual treatment effects from responses and a predictive biomarker in a parallel group RCT. 2014 [cited 2015 Jan 04]. Available from: http://epub.ub.uni-muenchen.de/22207/.
29. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Stat Med. 2013; 32: 2262–2277.
30. Harrell FE. Regression Modeling Strategies. New York: Springer; 2001.
31. Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. Stat Med. 2013; 32: 3788–3803.
32. Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of

power for several methods of analysis. Stat Med. 2014; 33: 4695–4708.
33. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother. 2004; 48: 2787–2792.
34. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005; 92: 965–970.
35. Royston P, Sauerbrei W, Ritchie A. Is treatment with interferon-alpha effective in all patients with metastatic renal carcinoma? A new approach to the investigation of interactions. Br J Cancer. 2004; 90: 794–799.
36. European Medicines Agency. Guidelines on the investigation of subgroups in confirmatory clinical trials. Pages 1–20, 2014.
37. Matsui S, Simon R , Qu P, Shaughnessy JD Jr, Barloqie B, Crowley J. Developping and validating genomic signatures in randomized clinical trials in predictive medicine. Clin Cancer Res. 2012; 18: 6065–6073.
38. Engelhardt A, Shen YM, Mansmann U. Constructing an ROC Curve to Assess a Treatment-Predictive Continuous Biomarker. Stud Health Technol Inform. 2016; 228: 745–749.