Graphical presentation of patient-treatment interaction elucidated by continuous biomarkers

Yu-Ming Shen

München 2016

Graphical presentation of patient-treatment interaction elucidated by continuous biomarkers

Yu-Ming Shen

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)

Aus dem Institut für Medizinische Informationsverarbeitung,
Biometrie und Epidemiologie
an der Medizinischen Fakultät
der Ludwig–Maximilians–Universität München

vorgelegt von
Yu-Ming Shen
aus Pingtung, Taiwan

München, den 17 Oct. 2016

Erstgutachter: Prof. Dr. rer. nat. Ulrich Mansmann
Zweitgutachter: Prof. Dr. rer. nat. Anne-Laure Boulesteix
Drittgutachter: Priv.-Doz. Dr. med. Sebastian Stintzing
Tag der mündlichen Prüfung: 17 Oct. 2016

Contents

List of Figures

List of Tables

Abstract

The translation of complex statistical results into clinical practice on the roles of continuous biomarkers in patient-treatment interactions is greatly aided by clear graphical presentation. To combat the current lack of comprehensive reviews or adequate guides on graphical presentation within this topic, our study formulates guiding principles for continuous biomarker in patient treatment interaction (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.

Several criteria for a good CBPTI plot are derived in this study, including general principles of visual display, appropriate quantification of statistical uncertainty, use of units presenting absolute outcome measures, correct display of benchmarks, and information content for medical decision-making. We examined a representative sample of biostatistics and clinical reports on randomized controlled trials with parallel-group design, based on papers published in four major biostatistics journals and two clinical trial methodology journals from the years 2000-2014, and six major clinical journals from 2013-2014. Each CBPTI plot found was assessed for appropriateness of its presentation and clinical utility.

In the systematic review, a total of seven methodological papers and five clinical reports used CBPTI plots which we categorized into four types: distinguishing the outcome effect for each treatment group, showing outcome difference among treatment groups (by either partitioning all individuals into subpopulations or modelling the functional form of the interaction), evaluating the proportion of population impact of the biomarker, and showing the classification accuracy of the biomarker. The current practice of utilizing CBPTI plots in clinical reports suffers from several poor practices: confusing or unclear labelling in the plot, 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 for medical decision-making.

There is considerable scope for improvement in the graphical representation of CBPTI in clinical reports. The existing statistical toolbox is not fully translated into clinical research and also needs improvement. The current challenge is to develop instruments for high-quality graphical plots which can not only convey quantitative concepts to readers with limited statistical knowledge when sophisticated statistical algorithms are undertaken, but also facilitate medical decision-making.

Chapter 1

Introduction

Consider a trial in which individuals are randomized to either standard or experimental treatment. The primary aim of conducting such a trial is usually the estimation of the overall treatment effect. Given a weak overall effect and the effort and cost involved in this 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 cetuximab as the treatment for metastatic colorectal cancer patients [38].

To represent treatment-biomarker interaction, graphical presentations are often used. For a set of dichotomous/categorical treatment-selection biomarkers, a modified forest plot can present heterogeneity of treatment effects within subpopulations [10]. In the case of continuous biomarkers, this tool cannot be applied, except if the continuous biomarker is categorized. But categorization destroys information [3] and creates several statistical problems: (1) the question of whether categorization of a continuous biomarker preserves randomization [11], (2) statistical multiplicity issues due to selecting an optimal cut-off value [2], and (3) instability of the statistical significance of the treatment-biomarker interaction depending on the number and positions of cut-off values [27]. Therefore, a first approach should involve analyzing the continuous biomarker 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 [27].

Tools to graphically present differential effects between a continuous biomarker and specific treatments exist in the literature, but have not been developed systematically [28]. The most popular approaches are treatment-effect plots [27] 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 [27]. 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] proposed STEPP, which uses pseudo-spline functions for exploring treatment-effect

heterogeneity across the range of a continuous biomarker (in terms of hazard ratios or differences in survival probabilities). 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.

For example, Figure 1.1 is an example of STEPP and 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 expression. Estrogen receptor 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 under the combination, a value above 1 indicates longer survival 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.

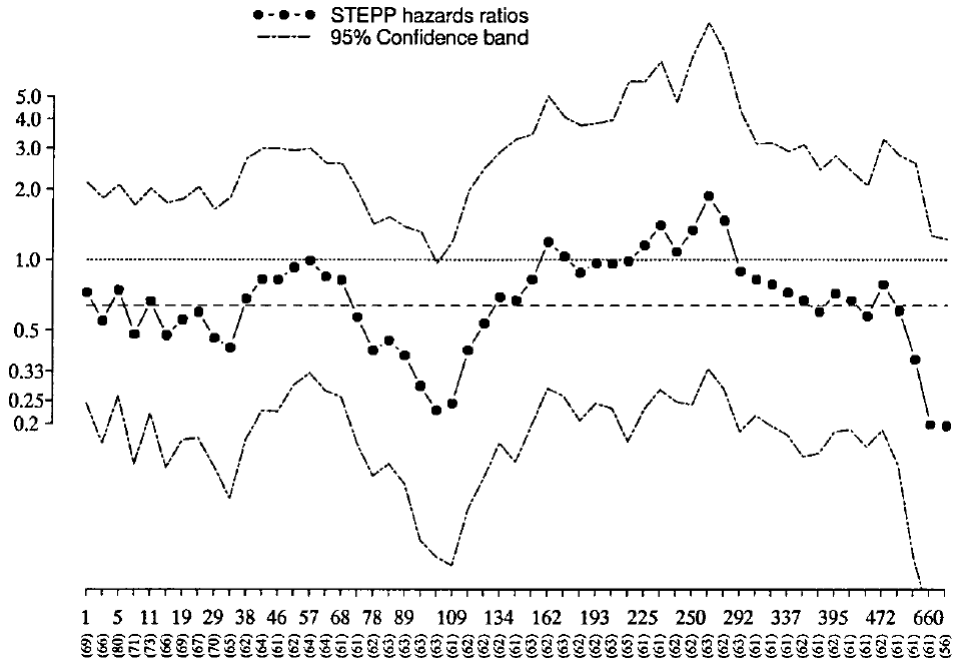

Figure 1.1: STEPP (sliding-window analysis) for International Breast Cancer Study Group Trial VII data according to estrogen receptor expression values ($n_1 = 55, n_2 = 60$). (Bonetti et al. Stat Med Oct 15, 2000 p2601) [8].

How helpful is this plot for the clinician? Is there convincing evidence for biomarker-

treatment interaction or treatment effect heterogeneity in the trial population? Can the clinician determine a subgroup of patients who profit from the combination therapy by measuring the ER expression? Can the graph help to derive individual therapeutic decisions? What are the guiding principles for constructing a continuous biomarker in patient treatment interaction (CBPTI) plot which is helpful in answering these 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 [23, 24, 37]. However, although general principles are available on what constitutes good practice in representing 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 [25, 36], the types of confidence intervals/bands to quantify statistical uncertainty [9], improving direct interpretation by adding a benchmark line [10], using the same scale to facilitate comparison of candidate biomarkers [16, 33], and focusing on answering key clinical questions about the proportion of impacted patients given the use of various biomarker measures to select treatment [33, 18]. 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 1.1 fulfills most of these criteria. But can it help to answer the above-mentioned key clinical questions?

The aim of our study is to formulate guiding principles for CBPTI plots. In order to understand the current practice, we review how CBPTI plots are used in clinical research (it remains unclear the extent to which 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 vignette which applies our ideas in a very simple setting.

Chapter 2

Methods

2.1 Criteria for a good CBPTI plot

Previous literature has discussed several aspects relating to the quality of the figure which also apply to CBPTI plots:

2.1.1 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 a wrong 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 [9]. 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.

2.1.2 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 [39]. 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 [25]. 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 [36]. An example of statin therapy reducing the risk of major coronary events is given in their report [36]. 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.

2.1.3 Benchmarks

The issue of benchmarking is of particular interest. 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 [10] 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 overall treatment effect answers the key question of assessing heterogeneity between subgroups.

2.1.4 Informative for medical decision-making

A good CBPTI plot should provide informative content for medical decision-making. What kind of information in a CBPTI plot is clinically helpful? Often authors present CBPTI plots by calculating the treatment effect across the range of the biomarker, and may add the

p-value of an interaction test to tell readers whether there is heterogeneity of treatment effect across the biomarker values. However, this approach is inadequate since the plot cannot help physicians select treatments. Janes et al propose several key functions of a good CBPTI plot [18]: to help clinicians choose one treatment over others for patients on the basis of a biomarker, to tell what proportion of a population would have a good response due to the treatment selection strategy, what proportion of patients would have treatment changes after biomarker measurement, or which biomarker is best if several candidates exist. Nevertheless, there are still many clinical questions needing to be answered. For example, we may be interested in the conditional probability given a biomarker range that the experimental treatment is better than the standard treatment. The outcome measures should be modelled as a function of the biomarker instead of the outcome measure itself since of interest is prediction of treatment selection for individuals, conditioned on the biomarker value. A good CBPTI plot should guide clinicians, their patients, and health policy makers to make good decisions in practice.

2.1.5 Summary

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 [23, 24, 37]. Here, we list critical guiding principles for CBPTI plots, in particular for future practice.

- To ensure the high quality of the plot, one needs to carefully take into consideration the principles of visual display suggested by the experts.
- Display appropriate measures of statistical uncertainty, either a pointwise confidence interval or simultaneous confidence band.
- Use absolute units for presenting outcome measures.
- 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.
- A good plot should be intrinsically informative for medical decision-making in a clinical setting.

2.2 Literature review

The literature review in our study was conducted using biostatistics journals, medical journals and journals for clinical trials methodology. 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 key words. 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 was not used [4]. The selection of studies based on important criteria relevant for our systematic review – how they present CBPTI plots and whether or not biomarkers are on a continuous scale – is not readily support by PubMed/MEDLINE. Therefore, we did a hand search in selected journals to include eligible reports.

The literature review in our study consisted of two parts: a review of methodologies in biostatistics and clinical trial journals, and a review of clinical applications in medical journals. For the former, four biostatistics and two clinical trial journals were selected: *Biometrics*, *Biostatistics*, *Statistics in Medicine*, *BMC Medical Research Methodology*, *Clinical Trials*, and *Trials*. The first four journals are the main biostatistics sources for topics relevant to the application of statistics to clinical trials and aim to enhance communication between statisticians and medical researchers. The last two journals publish articles mainly on general trial methodology. The systematic 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 [12] and Bonetti et al [8] in 2000. 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: *The New England Journal of Medicine*, *The Lancet*, *The Journal of the American Medical Association*, *Annuals of Internal Medicine*, *The Lancet Oncology* and *Journal of Clinical Oncology*. The first four are flagships for reporting general clinical innovations; the last two are flagships for innovation in clinical oncology, a very active field of personalized medicine and biomarker-driven treatment decisions. A previous report indicated that articles concerning treatment-selection biomarkers are far more likely to appear in high impact journals (24.7% of all such articles) than low impact journals (11.6%) [35]. 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 clinical and methodological 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.

Chapter 3

Results

In the biostatistics and clinical trial methodology journals, there were five reports relevant to CBPTI plots during 2000-2014, including two in Biometrics [16, 33], two in Statistics in Medicine [27, 8], one in Biostatistics [9], and none in BMC Medical Research Methodology, Clinical Trials or Trials. In addition to these five reports, two newly developed CBPTI plots were found through cross-referencing: marker-by-treatment predictiveness curves [18] and risk curves [17]. 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 [5, 41, 22, 21, 40] 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 [5, 41, 22, 21]. 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:

3.1 Evaluation for a CBPTI plot

3.1.1 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 3.1 [5] and Figure 3.2 [41] are examples of interaction plots found in medical journals.

The aim of Figure 3.1 is to explore the influence of age on the treatment effect, which is measured as the change in symptom Distress Score (SDS)-15 during the study [5]. The methodology behind the plot was based on nonparametric smoothing techniques. For

details see Berry et al. [5]. Figure 3.1 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 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 below the curve corresponding to the control. Both curves are nearly identical for age below 50. Figure 3.1 is not informative for clinicians in therapeutic decision-making. Although the authors present individual outcomes by dots, the distributions are identical. The dots provide no information on population impact when using age as a biomarker. For prediction of individual treatment selection, outcome should be specified as a function of the biomarker. The use of change score as an outcome measure is also problematic since change score is affected by regression to the mean. A Bland-Altman plot is recommended for this example [7]. The lack of uncertainty is a major limitation of Figure 3.1.

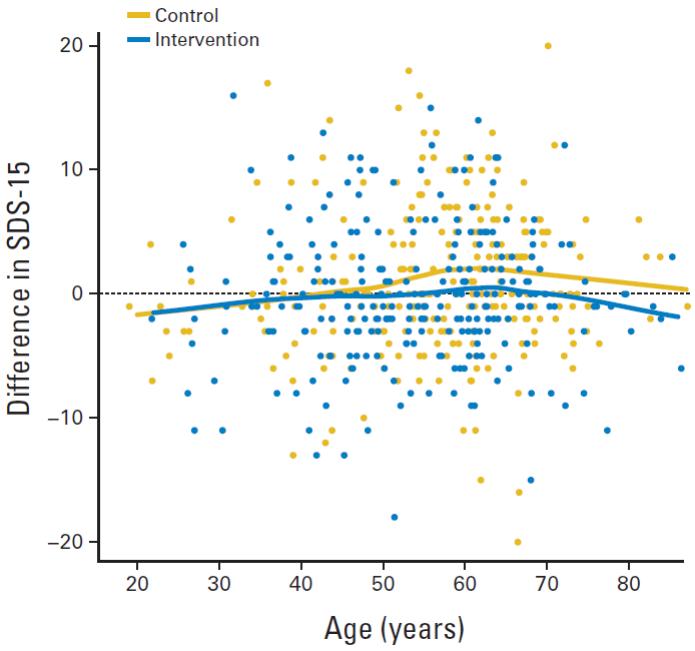

Figure 3.1: Effect of age on Symptom Distress Scale-15 (SDS-15) score change between baseline and end of study. (Berry et al. JCO Jan 20, 2014 p204) [5]

Figure 3.2 is another example of an interaction plot but extends to evaluate treatment-biomarker interaction under multi-subsets [41]. 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. Each curve is easily estimated by adding interaction terms to the Cox proportional hazards model. On the

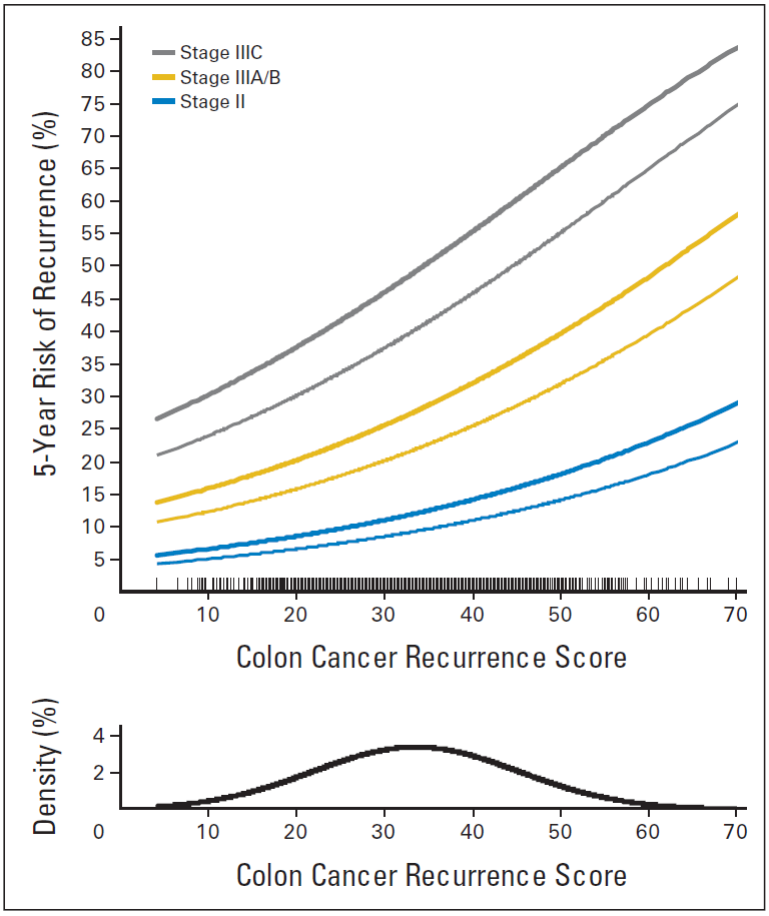

Figure 3.2: Relationship between the continuous Recurrence Score (RS) and 5-year recurrence risk by stage and treatment in the National Surgical Adjuvant Breast and Bowel Project C-07. Thick lines represent fluorouracil (FU)-treated patients, thin lines represent FU + oxaliplatin-treated patients. Blue, 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 p4515) [41]

basis of the principles of visual display, Figure 3.2 shows two 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. However, a total of six lines in a plot make it difficult for the eyes 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. There is no information to help clinicians select treatment. 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 have the distribution of the CCRS within each stage (assuming a randomization of the treatment). Again there are no confidence intervals to quantify uncertainty.

3.1.2 Showing outcome difference between treatment groups

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 the introduction part. In Figure 1.1 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 estrogen receptor expression. Subpopulation sizes for values of the biomarker are added alongside the x-axis. However, this is not very informative for medical-decision making since subpopulations are allowed to overlap in order to increase the number of subjects who contribute to each point estimate. Figure 3.3 is a clinical application of STEPP from a medical journal [22] and has similar limitations to those of Figure 1.1.

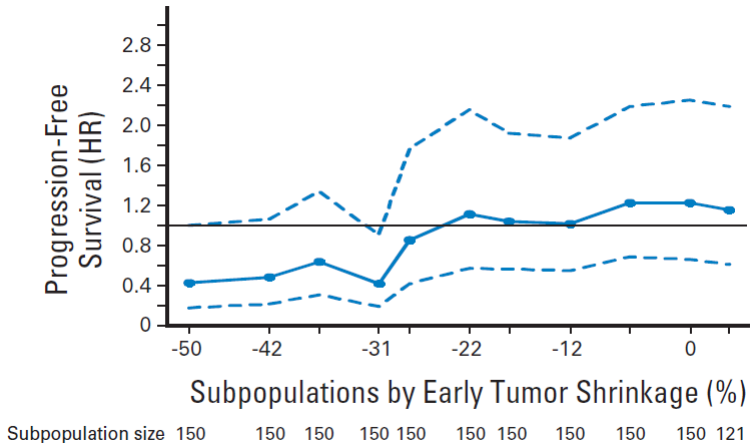

Figure 3.3: Sliding-window subpopulation treatment effect pattern plot analysis of the treatment effect of adding cetuximab to chemotherapy in patients with KRAS wild-type as measured by hazard ratios (HRs) for progression-free survival (chemotherapy plus cetuximab v chemotherapy alone. HR values < 1 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 p3770) [22]

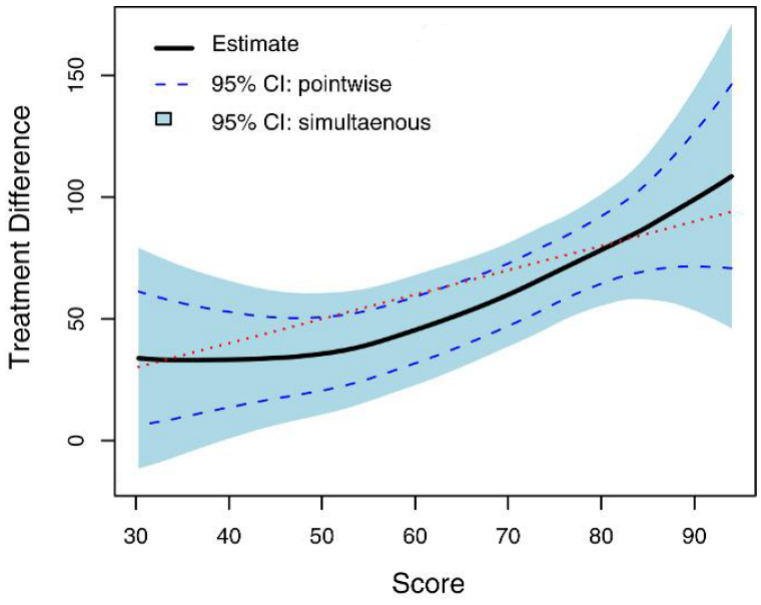

Figure 3.4: Estimated treatment differences (thick curve), 3-drug combo minus 2-drug combo, with respect to week 24 CD4 changes over the score and the corresponding 95% pointwise (dashed curve) and simultaneous (shaded region) confidence intervals. (Cai et al. Biostatistics Apr, 2011 p277) [9]

Cai et al proposed a more advanced and formally precise approach to presenting outcome difference between treatment groups [9]. 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 3.4 displays a shaded region and dashed curves to identify two types of uncertainty estimates [9]. 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 would be recommended for the plot to improve direct interpretation. The plot adds no crucial information which would influence a therapeutic decision, since the 3-drug combination always performs better than the 2-drug combination over the scores.

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 3.1. 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 $\Delta(score) = \alpha + \beta \cdot age + \gamma \cdot treat + \delta \cdot treat \cdot age$, and presents the contrast line as $f(age) = \gamma + \delta \cdot age$. 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 [27]. 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 calculating a difference. The difference could only be given if for single patient their counterfactual outcomes under the second treatment could be known. Recent approaches have tried to estimate individual treatment effects within a counterfactual framework [16, 19]. 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 [27, 6, 14]. Royston and Sauerbrei proposed the use of a power transformation, termed "fractional polynomial", in modeling the functional form of a continuous biomarker [27]. 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.

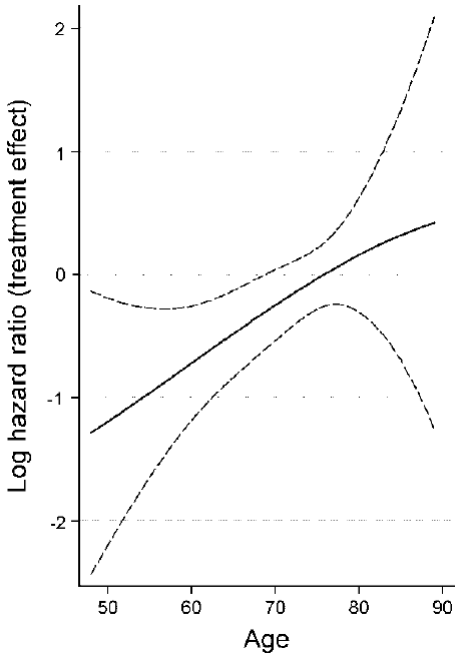

Figure 3.5: Prostate cancer data: treatment × age interaction: the effect of treatment by age, with 95% pointwise confidence interval. Functions were estimated in multivariable adjustment models and fitted using FP2 functions with powers (3; 3). (Royston et al. Stat Med Aug 30, 2004 p2516) [27]

Figure 3.5 is an example of a CBPTI plot based on their approach, named "treatment-effect plot" [27]. 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 for correct and fast interpretation.

Harrell proposed an approach of employing restricted cubic splines in modelling the functional form of a continuous covariate [14] 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. With this model fit plots simultaneous confidence bands for the treatment difference. Other proposed existing spline methods could be employed in modelling interaction between a continuous biomarker and treatment as well. Figure 3.6 is a clinical example from a medical review [21]. 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).

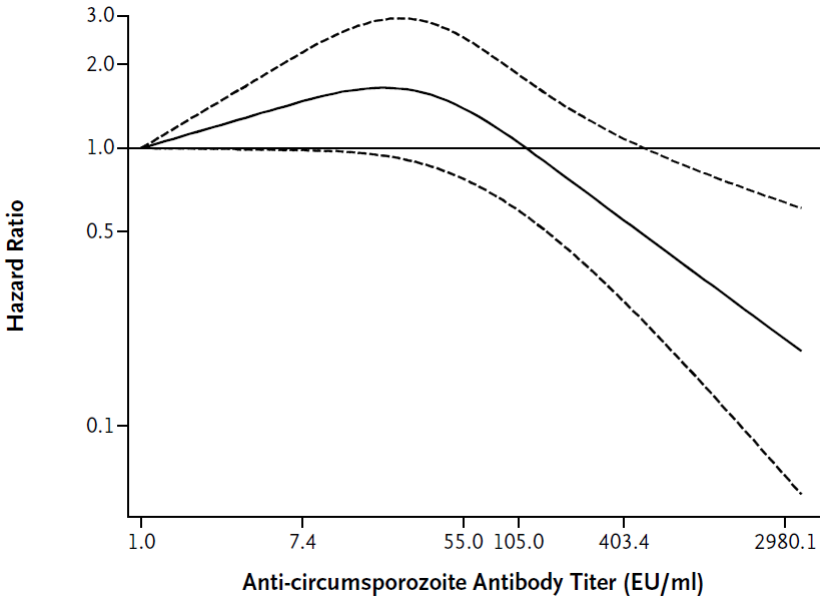

Figure 3.6: The association between imputed anti-circumsporozoite antibody titers and the hazard ratio for clinical malaria episodes among children who received the RTS,S/AS01E vaccine, according to a Cox regression model with cubic splines and with a baseline titer of 1.0 enzyme-linked immunosorbent 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 p1119) [21]

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 [29, 30].

Hazard ratios are presented in Figure 1.1,3.3, 3.5, 3.6, and may be not good way to interpret quantity. We can transform a hazard ratio into risk probability [34].

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

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 [13], 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 3.7. The plot shows that the proportion 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 [31]. A horizontal line for a benchmark can be drawn for the decision of the threshold. A threshold where the solid curve reaches the point of 50% of the population is recommended.

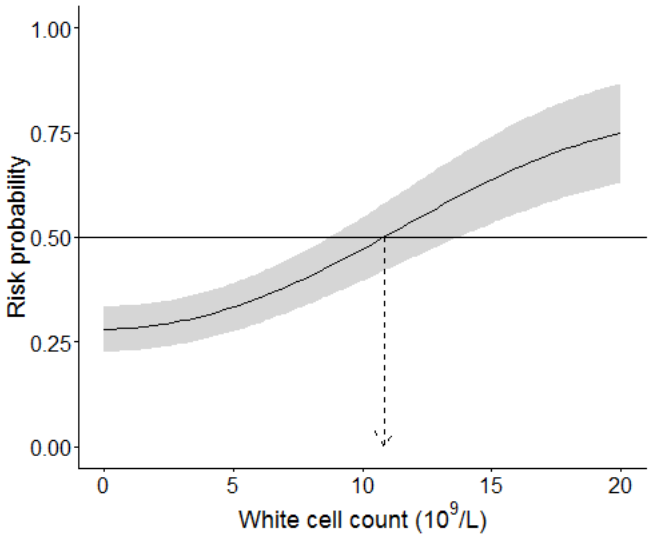

Figure 3.7: Risk probability for a subject with interferon-alpha treatment experiencing the event before a patient with medroxyprogesterone acetate treatment. (The dataset was obtained from a Medical Research Council RE01 phase III randomized controlled trial [31])

3.1.3 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 [33, 18]. The methodologies relevant to the proportion of population impact of the biomarker include the selection impact curve [33], the marker-by-treatment predictiveness curve [18], and the risk curve [17].

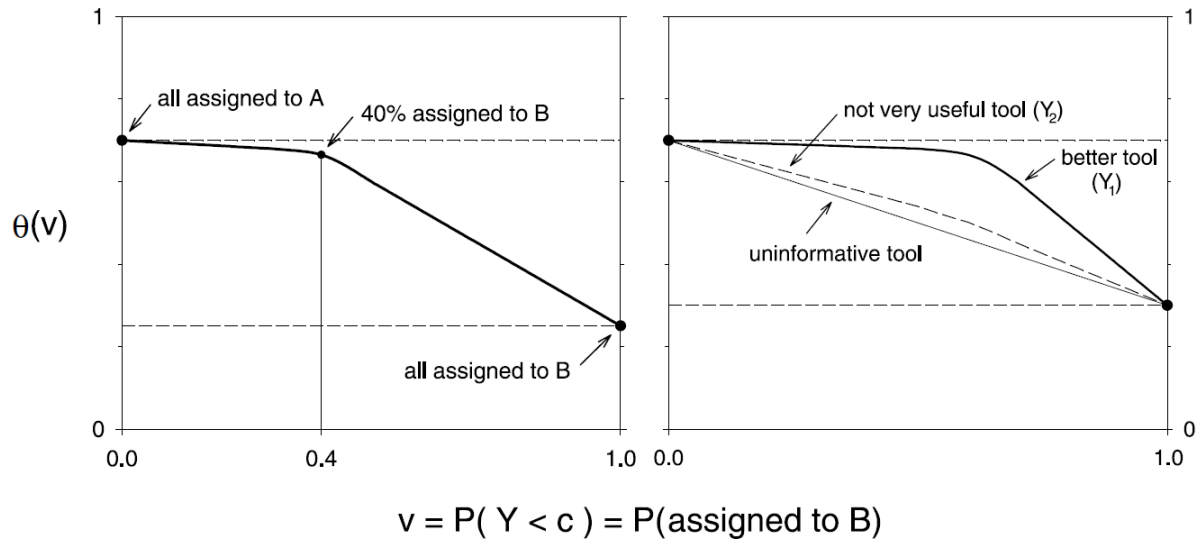

Figure 3.8: A schematic diagram of the selection impact (SI) curve, $\theta(v)$ = the population response rate $P\{D = 1|(Y > c, T = 1)or(Y < c, T > 0)\}$. (Song et al. Biometrics Dec, 2004 p875) [33]

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 3.8 [33]. For health-decision makers, it is helpful to identify the proportion of population that are more likely to beneift 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 $(1, max\{\text{response rate}\})$. 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, due to a result of the paper being published in a biostatistics journal and the readers of which being primarily statisticians. Confidence bands displaying statistical uncertainty are encouraged to ease comparability.

The other innovative CBPTI plots proposed by the same research group, the marker-by-treatment predictiveness curve (Figure 3.9) [18] and the risk curve (Figure 3.10) [17], 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 absolute units. The vertical lines for the threshold of treatment change allow seeing which treatment delivers the best outcome given the biomarker mea-

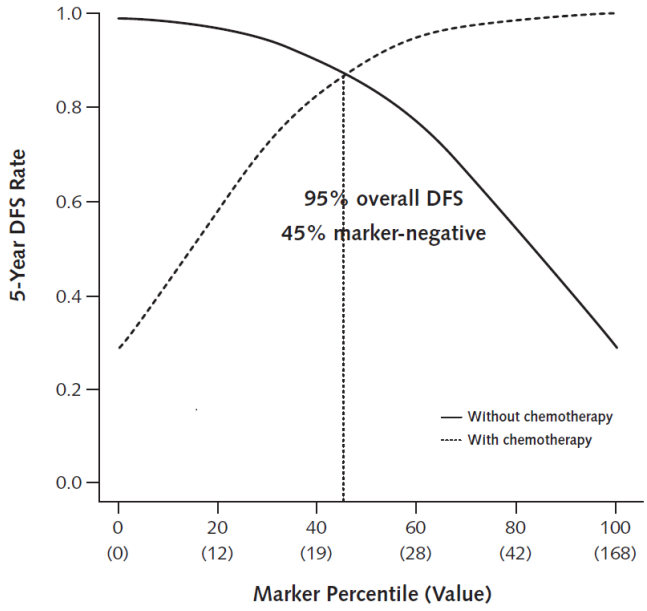

Figure 3.9: The 5-year disease-free survival (DFS) rate plotted as a function of marker percentile, with raw marker values shown in parentheses. The overall DFS rate with use of the marker for guiding treatment is shown, as well as the percentage of women who have higher DFS rates with tamoxifen alone (marker-negative). (Janes et al. Ann Intern Med Feb 15, 2011 p255) [18]

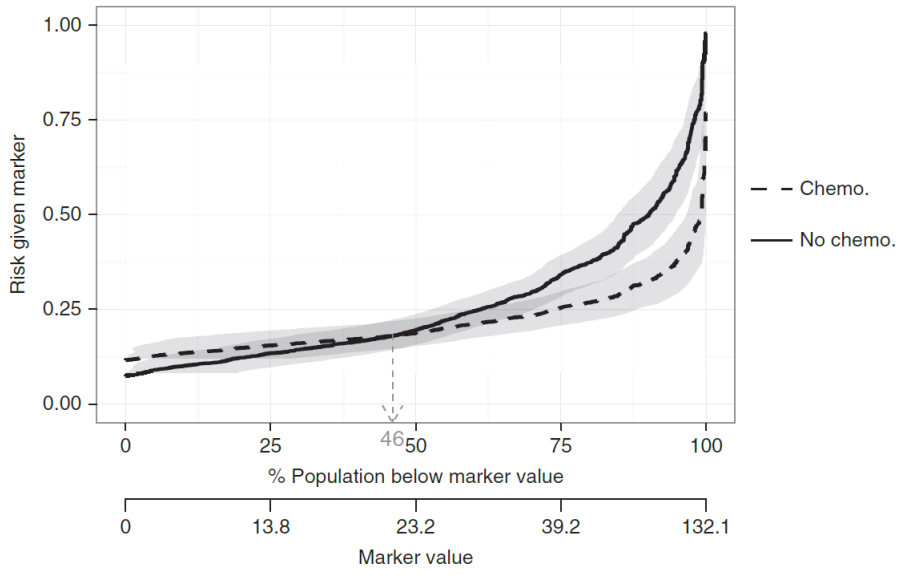

Figure 3.10: Risk of 5-year breast cancer recurrence or death as a function of treatment assignment and marker percentile, for the Oncotype-DX-like marker. Horizontal pointwise 95% confidence intervals (CIs) are shown. Forty-six percent of women have negative treatment effects according to the Oncotype-DX-like marker; these women can avoid adjuvant chemotherapy. (Janes et al. Int J Biostat 2014 p102) [17]

surement. For Figure 3.9, the confidence intervals for statistical uncertainty are missing. Note there is no need for a benchmark line in this type of CPBTI plot because we are interested in how the two curves deviate from each other. A larger variation implies a better performance of the biomarker in differentiating the treatment effect.

3.1.4 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 [16]. 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.

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 (treatment-effective individual), and the false positive fraction, (FPF) defined as the probability of incorrectly classifying a bad responder (treatment-ineffective individual) 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 3.11 comes from Huang's approach but similar to our idea for two candidate treatment-selection biomarkers [16]. 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 value (PPV) and negative predictive value (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

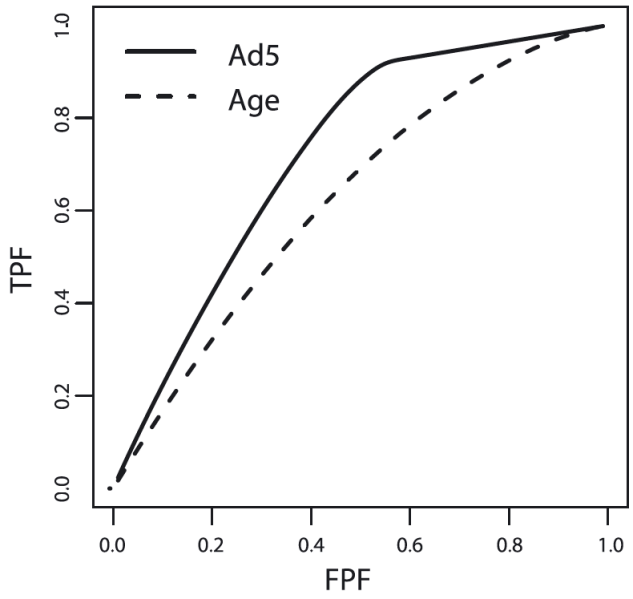

Figure 3.11: Plot of ROC curves for classifying a subject into treatment-effective or treatment-ineffective groups. (Huang et al. Biometrics Sept, 2012 p694) [16]

proportions of good responders and bad responders in biomarker measurement that are above cut-off value or below 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 and bad responders 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 section 3.2.5 *Prediction curve*.

3.2 R cbpti vignette

R "`cbpti`" vignette was developed to explore the interaction between a linear or nonlinear (fractional polynomial) biomarker and treatment group. The package can be implemented in linear, logistic, and Cox regression. In the following we will describe main parts of the R comments for cbpti by using an example from Medical Research Council RE01 phase III randomized controlled trial [31].

In the dataset, 347 patients with metastatic renal carcinoma were randomized to either to interferon-alpha treatment (IFA, $n = 172$) or medroxyprogesterone acetate treatment (MPA, $n = 175$) at 31 centers in the UK between 1992 and 1997. The data consists of ID, treatment group (0=MPA, IFA=1), patient's age at baseline, patient's white blood cell count, censoring indicator (0=censored and 1=event), and overall survival time (days). Of these patients, 25 patients were censored and their survival times were imputed [26].

Therefore, the censoring indicators for all patients were recorded as "1". The data can be loaded from the package by applying the function `data()`, specifying the package "`cbpti`" to be chosen.

```
R> data(MRC,package="cbpti")
```

The users can use the function `str()` to display the internal structure of dataset `MRC`. Here, the treatment outcome is overall survival and survival time was converted from days to months. The patient's white blood cell count ($10^9/L$) was used as a biomarker.

```
R> str(MRC)
'data.frame':  347 obs. of  6 variables:
 $ ID      : num  1037 1074 1149 1324 1046 ...
 $ trt     : num  1 1 1 1 1 0 1 1 1 1 ...
 $ age     : num  67.3 42.1 65.4 60.5 67.8 ...
 $ wcc     : num  6.4 5.1 9.2 7.3 6.6 ...
 $ censdead: num  1 1 1 1 1 1 1 1 1 1 ...
 $ survtime: int  1419 2254 271 843 611 473 1098 1057 1051 233 ...
```

```
R> MRC$censtime <- round(MRC$survtime/30, digits=2)
```

First, the users can use the function `bm.form` to find the best fitted fractional polynomial object for a continuous biomarker and use the function `print()` to display the results.

```
R> form <- bm.form(Surv(censtime, censdead) ~ fp(wcc, df=4, select=NA,
+          scale=TRUE), data=MRC, arm="trt", family=cox, method="efron")
```

```
R> print(form)
```

```
Call:
bm.form(formula = Surv(censtime, censdead) ~ fp(wcc, df = 4, select = NA, scale =
        TRUE), data = MRC, arm = "trt", family = cox, method = "efron")
```

```
Arm = 0
$coef.0
                  coef exp(coef)  se(coef)        z     Pr(>|z|)
I((wcc/10)^1) 0.7289966  2.072999 0.2186012 3.334824 0.0008535354
Concordance: 0.576 Standard error: 0.026
```

```
Arm = 1
$coef.1
                                    coef exp(coef)   se(coef)         z     Pr(>|z|)
I((wcc/10)^3)                  0.8999309 2.4594331 0.12853702  7.001336 2.535305e-12
I((wcc/10)^3 * log((wcc/10))) -0.6538387 0.5200456 0.09892509 -6.609433 3.857947e-11
Concordance: 0.647 Standard error: 0.026
```

where `fp` denotes a special term for fitting fractional polynomial. There are several argu-

ments in 'fp'. The 'df' is degrees of freedom of the fractional polynomial model. The df=4 and df=2 denote the maximum permitted degree m=2 (default) and m=1 in fractional polynomial model, respectively. The df=1 is linear fractional polynomial model. The argument `select` sets the variable selection level for the input variable. The argument `alpha` sets the fractional polynomial selection level for the input variable. The argument `scale` specify whether to use pre-transformation scaling to avoid numerical problems (default=TRUE). The argument `arm` is treatment group assignments. Error occurs if not equal to two treatment arms are found in this variable.
The results show that the functions of white blood cell count for MPA treatment and IFA treatment are

$$\beta_1 \times \Big(\frac{wcc}{10}\Big)$$

and

$$\beta_1 \times \Big(\frac{wcc}{10}\Big)^3 + \beta_2 \times \Big(\frac{wcc}{10}\Big)^3 \times log\Big(\frac{wcc}{10}\Big),$$

respectively.

In the following, (1) interaction plot, (2) contrast plot, (3) proportion of unfavorable treatment effect plot, (4) ROC curve, and (5) prediction curve were produced presenting the relationship between treatment group and white blood cell count. In the R vignette, pointwise confidence intervals was calculated instead of simultaneous confidence bands since it is much easier to calculate than simultaneous confidence bands and commonly used in clinical reports. Moreover, previous study has shown that pointwise confidence intervals work well in practice [15]. For the sake of simplicity, here assumes that the outcome measurements of the two treatment groups are not correlated and the covariance is thus assumed to be zero. We give the functions written in R accompanied by the produced plots. Note that the functions were written and tested and the plots produced with R version 3.2.2.

3.2.1 Interaction plot

The users can produce the interaction plots with 95% pointwise confidence intervals by applying the function `bmplot.interaction()`. The function draws fitted a regression line for each treatment arm. The dots show the risk of death of population distribution relative to white blood cell count. The argument `obs.time` specifies the length of observation for survival data. The following example is interaction plot for the white blood cell count. The function `autoplot()` and `print()` are generic functions for plotting an object and printing the results, respectively.

```
R> p1 <- bmplot.interaction(Surv(censtime, censdead) ~ fp(wcc, df=4, select=NA,
+        scale=TRUE), data=MRC, family=cox, arm="trt", obs.time=6,
+        conf.level=0.95)
```

```
R> print(p1)
```

```
*.0 indicates arm = 0
*.1 indicates arm = 1
```

```
    wcc.interval     lwr.0     fit.0     upr.0      lwr.1        fit.1        upr.1
1       3.100000 0.6369106 0.5302407 0.4235708  0.3955423 3.434533e-01   0.29136437
2       3.250578 0.6388644 0.5341418 0.4294193  0.3972799 3.452587e-01   0.29323743
3       3.401156 0.6408590 0.5380531 0.4352473  0.3991536 3.472049e-01   0.29525635
4       3.551734 0.6428947 0.5419742 0.4410537  0.4011672 3.492962e-01   0.29742520
5       3.702312 0.6449716 0.5459047 0.4468378  0.4033248 3.515364e-01   0.29974806
6       3.852890 0.6470901 0.5498443 0.4525986  0.4056302 3.539296e-01   0.30222898
7       4.003468 0.6492502 0.5537927 0.4583351  0.4080875 3.564797e-01   0.30487199
8       4.154046 0.6514523 0.5577494 0.4640464  0.4107006 3.591908e-01   0.30768113
9       4.304624 0.6536968 0.5617141 0.4697314  0.4134734 3.620669e-01   0.31066045
10      4.455202 0.6559838 0.5656864 0.4753891  0.4164100 3.651120e-01   0.31381399
:
(Omit)
```

```
R> autoplot(p1, xlim=c(0,20), ylim=c(0,1.25), xlab=expression(paste("White blood
+  cell count ", "("*10^{9})*"/L)"), ylab="Risk of death by 6 months", title="",
+  background=theme_classic()
```

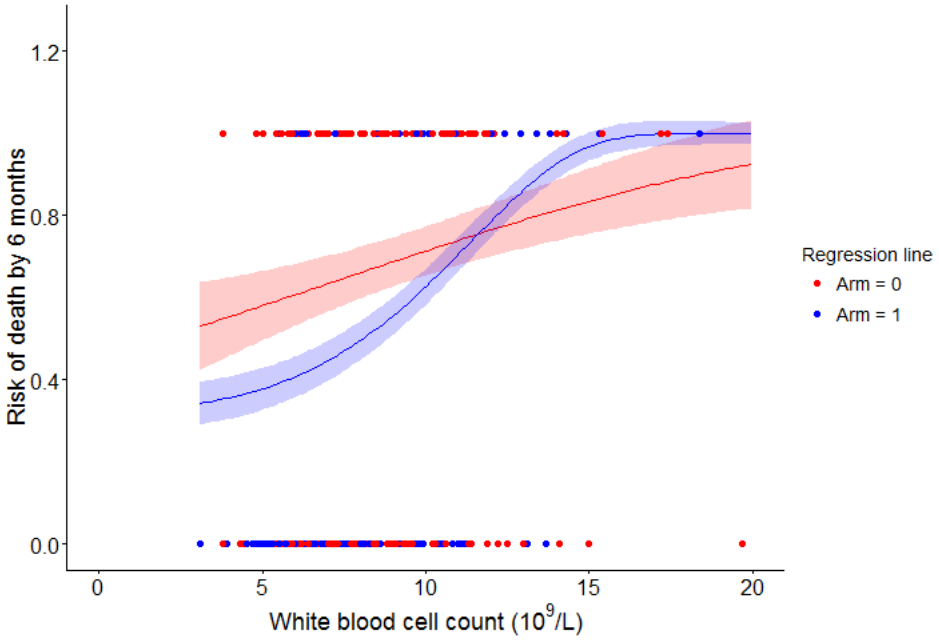

Figure 3.12: R cbpti vignette: Interaction plot

The interaction plot shows that the IFA treatment is only favorable for the subjects whose white blood cell counts below 11 ($10^9/L$).

3.2.2 Contrast plot

The contrast plot showing the difference of treatment effects between treatment arms with 95% pointwise confidence interval is given by the function `bmplot.constrast()`. The horizontal benchmark line is calculated by the mean of predicted difference between treatment arms and displayed as a solid line in the contrast plot. The argument `ref` specifies the reference group. The following R codes are the contrast plots for the white blood cell count:

```
R> p2 <- bmplot.contrast(Surv(censtime, censdead) ~ fp(wcc, df=4, select=NA,
+        scale=TRUE), data=MRC, family=cox, arm="trt", ref="0", obs.time=6,
+        conf.level=0.95)
```

```
> print(p2)
```

```
    wcc.interval          diff     diff.se      diff.lwr     diff.upr
1       3.100000 -0.1867873517  0.06056546  -0.305493469 -0.068081234
2       3.250578 -0.1888831745  0.05965886  -0.305812399 -0.071953950
3       3.401156 -0.1908481767  0.05876798  -0.306031300 -0.075665053
4       3.551734 -0.1926780005  0.05789296  -0.306146116 -0.079209885
5       3.702312 -0.1943682980  0.05703396  -0.306152812 -0.082583784
6       3.852890 -0.1959147224  0.05619116  -0.306047368 -0.085782077
7       4.003468 -0.1973129215  0.05536472  -0.305825770 -0.088800074
8       4.154046 -0.1985585314  0.05455481  -0.305484003 -0.091633060
9       4.304624 -0.1996471720  0.05376164  -0.305018049 -0.094276295
10      4.455202 -0.2005744438  0.05298538  -0.304423882 -0.096725006
11      4.605781 -0.2013359256  0.05222623  -0.303697462 -0.098974389
12      4.756359 -0.2019271745  0.05148440  -0.302834738 -0.101019611
13      4.906937 -0.2023437257  0.05076008  -0.301831646 -0.102855805
14      5.057515 -0.2025810953  0.05005348  -0.300684109 -0.104478082
15      5.208093 -0.2026347836  0.04936481  -0.299388041 -0.105881527
16      5.358671 -0.2025002803  0.04869430  -0.297939348 -0.107061212
17      5.509249 -0.2021730711  0.04804214  -0.296333940 -0.108012203
18      5.659827 -0.2016486465  0.04740857  -0.294567727 -0.108729566
19      5.810405 -0.2009225117  0.04679378  -0.292636638 -0.109208386
20      5.960983 -0.1999901991  0.04619800  -0.290536623 -0.109443776
21      6.111561 -0.1988472824  0.04562144  -0.288263670 -0.109430895
22      6.262139 -0.1974893927  0.04506431  -0.285813818 -0.109164968
23      6.412717 -0.1959122374  0.04452680  -0.283183170 -0.108641305
24      6.563295 -0.1941116209  0.04400912  -0.280367914 -0.107855328
25      6.713873 -0.1920834680  0.04351145  -0.277364344 -0.106802592
:
(Omit)
```

```
R> autoplot(p2, xlab=expression(paste("White blood cell count ",
+  "("*10^{9})*"/L)"), ylab="Difference in Risk of death by 6 months",
+  xlim=c(0,20), ylim=c(-1,1), title="", background=theme_classic())
```

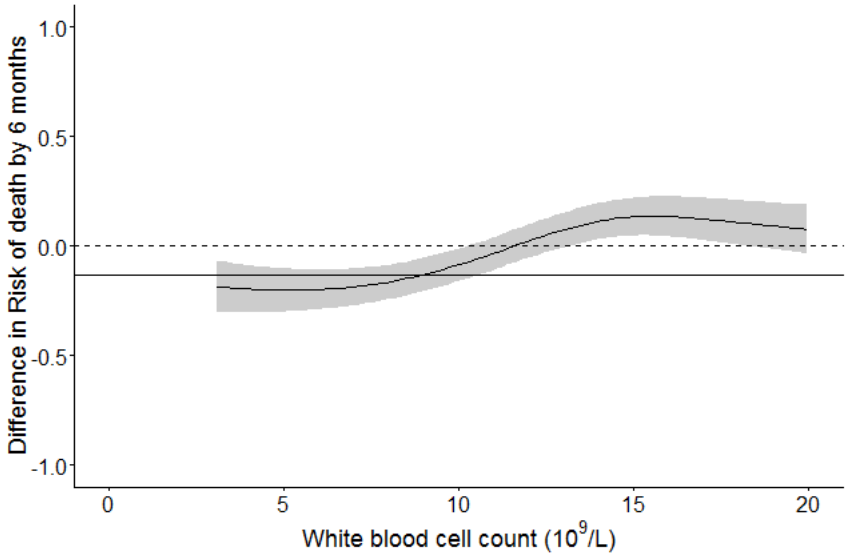

Figure 3.13: R cbpti vignette: Contrast plot

The contrast plot shows that the difference (IFA vs MPA) in risk of death by 6 months becomes higher with the increase of the white blood cell count.

3.2.3 Proportion of unfavorable treatment effect plot

The proportion of unfavorable treatment effect plot shows the estimated proportion of patients who would possibly suffer an unfavorable outcome from treatment and is given by the function `bmplot.proportion()`. A 95% pointwise confidence interval is also presented in the proportion of unfavorable treatment effect plot. The argument `ncoef` is the number of produced regression coefficients from the specified multivariate normal distribution and used for the estimation of confidence intervals. The following R codes are the proportion of unfavorable treatment effect plots for the white blood cell count:

```
R> p3 <- bmplot.proportion(Surv(censtime, censdead) ~ fp(wcc, df=4, select=NA,
+        scale=TRUE), data=MRC, family=cox, arm="trt", ref="0", obs.time=6,
+        conf.level=0.95, ncoef=1000)
```

```
R> print(p3)
```

```
    wcc.interval       pfit        lwr        upr
1       3.100000 0.39771477 0.37002992 0.42216171
2       3.250578 0.39663251 0.36756882 0.42231511
3       3.401156 0.39562186 0.36518035 0.42254460
4       3.551734 0.39468502 0.36286721 0.42285242
5       3.702312 0.39382418 0.36063208 0.42324080
6       3.852890 0.39304154 0.35847764 0.42371200
7       4.003468 0.39233931 0.35640656 0.42426829
8       4.154046 0.39171970 0.35442152 0.42491192
9       4.304624 0.39118495 0.35252520 0.42564521
10      4.455202 0.39073733 0.35072029 0.42647046
:
(Omit)
```

```
R> autoplot(p3, xlab=expression(paste("White blood cell count ",
+   " ("*10^{9})*"/L)"),ylab="Risk Probability in 6 months", xlim=c(0,20),
+   title="", background=theme_classic())
```

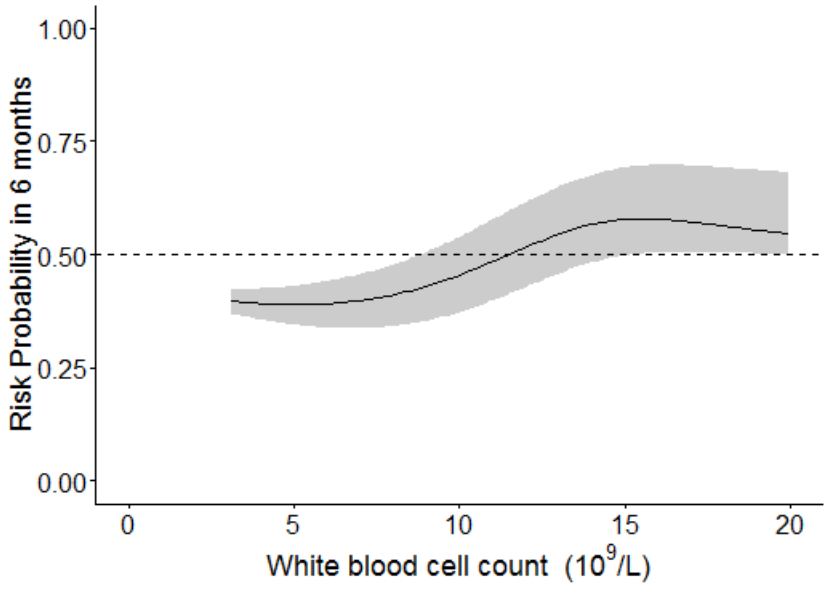

Figure 3.14: R cbpti vignette: Proportion of unfavorable treatment effect plot

In the package, there is no function to add a vertical line displaying the choice of threshold automatically. It is recommended to add a vertical line as a threshold where the curve reaches 0.5 in the proportion of unfavorable treatment effect plot; the implication is that the IFA treatment is suggested for the subjects whose white blood cell counts falls below 11 ($10^9/L$).

3.2.4 ROC curve

In regard to the ROC curve in the cbpti package, we do not employ Huang's methodology [16] since their approach uses ROC curves in the potential outcome framework under severe restriction. We generalize this idea to the standard two parallel armed RCT setting with continuous outcome and use a simple algorithm for estimation of the ROC.

The ROC displays true positive fractions and false positive fractions and can be plotted using the function `bmplot.ROC()`. The ROC analysis is constructed on a basis of contrast plot. For the treatment differences (Δ) between two groups given each biomarker value, a randomly generated sample of deltas from a normal distribution is created. A certain biomarker value is then specified as a cut-off point. If an individual's biomarker value is above this cut-off, this individual is considered a good responder, otherwise the individual is considered a bad responder. In this randomly generated population, we can categorize all individuals into four types: true positive, false positive, true negative and false negative individuals. The true positive fraction (TPF) is defined as the probability of correctly identifying a good responder. The false positive fraction (FPF) is defined as the probability of incorrectly classifying a bad responder as a good responder. Given different cut-off points, a ROC curve can be created and the area under the curve can be estimated. The argument `nperm` is the size of the generated random sample from the normal distribution and used for the improved estimation of true positive fraction and false positive fraction. The calculation of these estimates was done by applying the package ROCR [32].

```
R> p4 <- bmplot.roc(Surv(censtime, censdead) ~ fp(wcc, df=4, select=NA,
+        scale=TRUE), data=MRC, family=cox, arm="trt", ref="1", obs.time=6,
+        conf.level=0.95, ncoef=100, nperm=10)
```

```
R> print(p4)
```

```
       cutoff         fpr          tpr
1         Inf 0.000000000 0.0000000000
2   55.200001 0.004713523 0.0006094029
3   37.099998 0.007299490 0.0038566062
4   22.100002 0.009831081 0.0071930109
5   21.699999 0.012410257 0.0103854988
6   19.700001 0.014564900 0.0142053850
7   18.400000 0.016612600 0.0181093863
8   17.399998 0.018706023 0.0219886475
9   17.200001 0.022929961 0.0299341393
10  15.399998 0.025287961 0.0334842560
11  15.300000 0.027027027 0.0376485179
12  15.000001 0.029285347 0.0414412898
13  14.299999 0.034057587 0.0489707607
14  14.199999 0.038281015 0.0564895763
15  14.100001 0.040182064 0.0604088539
:
(Omit)
```

```
[1] "AUC=0.5525 (95% CI=0.5355-0.568)"
```

```
R> autoplot(p4, title="")
```

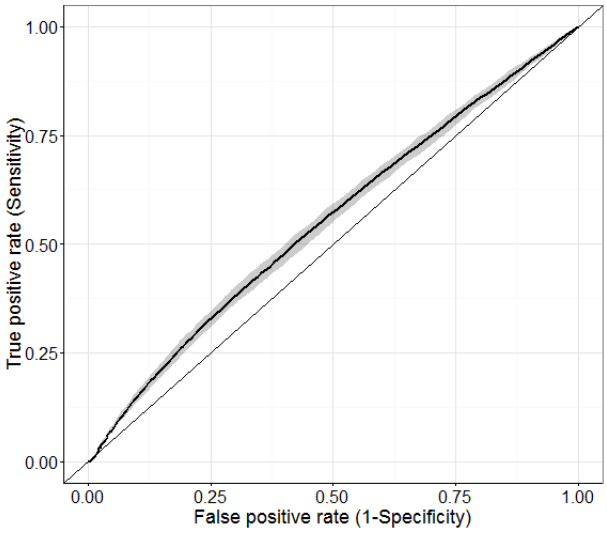

Figure 3.15: R cbpti vignette: ROC curve

The ROC curve may lie below the chance diagonal (implying the area under the curve (AUC) of the ROC curve < 0.5) depending on the direction of treatment effect across the range of biomarker value. In this case, we reversed the classifier by setting the other treatment group as reference and recalculate the AUC for the corrected classifier (corrected AUC=1-AUC), for easy interpretation. The interpretation of AUC for the white blood cell count is that a bad responder will have a higher white blood cell count than 55.2% (95%CI=53.5%-56.7%) of good responders. However, the result seems not to be so promising and points out that the white blood cell count does not discriminate between good responders and bad responders although we found a significant heterogeneity of treatment effect in interaction plot and contrast plot. For reporting CBPTI in clinical research, interaction plot and contrast plot are not sufficient for a good biomarker performance. ROC curve can be thought of good appraoch of presenting an intrinsic property of a biomarker.

3.2.5 Prediction curve

Since the true positive fraction and the false positive fraction are of no practical use in helping clinicians estimate the probability of good response in individual patients, here calculate the proportion of good responders and bad responders based on ROC analysis. Given each biomarker value, a randomly generated sample of deltas from a normal distribution is created. The proportion of good responders ($\Delta < 0$) and bad responders

($\Delta \geq 0$) can be estimated given each randomly generated sample for each biomarker value. The following R codes create the plot for the white blood cell count by using function `bmplot.prediction()` :

```
R> p5 <- bmplot.prediction(Surv(censtime, censdead) ~ fp(wcc, df=4, select=NA,
+        scale=TRUE), data=MRC, family=cox, arm="trt", ref="0", obs.time=6,
+        conf.level=0.95, ncoef=100, nperm=100)
```

```
R> print(p5)
```

```
       marker   bad.lwr     bad.p   bad.upr  good.lwr    good.p  good.upr
1    3.100000 0.3147500 0.3850000 0.4700000 0.5300000 0.6150000 0.6852500
2    3.800000 0.3415833 0.3966667 0.4601667 0.5398333 0.6033333 0.6584167
3    3.900000 0.2842500 0.3800000 0.4752500 0.5247500 0.6200000 0.7157500
4    4.300000 0.2947500 0.3800000 0.5052500 0.4947500 0.6200000 0.7052500
5    4.400000 0.3123750 0.3925000 0.4526250 0.5473750 0.6075000 0.6876250
6    4.500000 0.2895000 0.3900000 0.4852500 0.5147500 0.6100000 0.7105000
7    4.700000 0.3315833 0.3916667 0.4566667 0.5433333 0.6083333 0.6684167
8    4.800000 0.3348750 0.3887500 0.4601250 0.5398750 0.6112500 0.6651250
9    4.900000 0.3023750 0.3800000 0.4683750 0.5316250 0.6200000 0.6976250
10   5.000000 0.3286875 0.3825000 0.4501250 0.5498750 0.6175000 0.6713125
:
(Omit)
```

```
R> autoplot(p5, xlab=expression(paste("White blood cell count ",
+   "("*10^{9})*"/L)"), xlim=c(0,20), ylim=c(0,1), title="", background=
+   theme_classic())
```

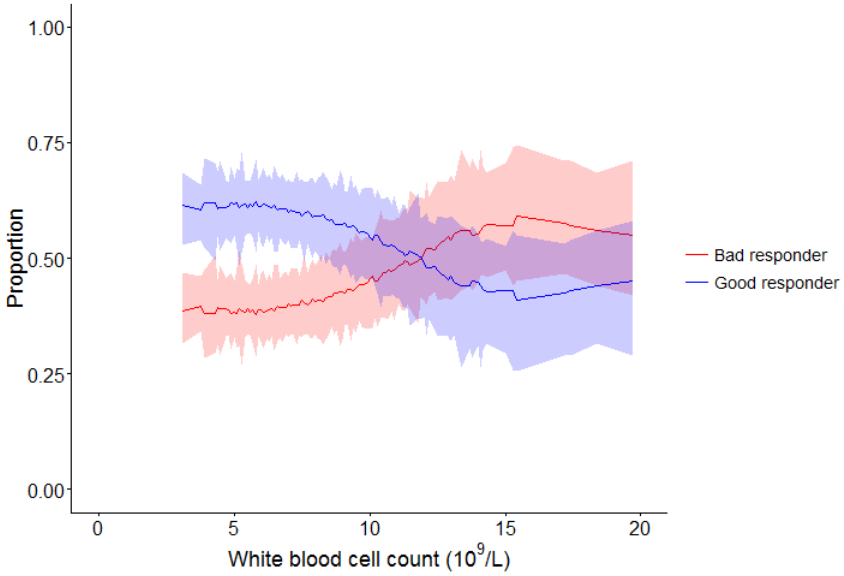

Figure 3.16: R cbpti vignette: Prediction curve

The blue and red colors denote the proportion of good responders and the proportion of bad responders under the IFA treatment, respectively. The proportion of good responders under the IFA treatment is higher than bad responders for the subjects whose white blood cell counts falls below 11 ($10^9/L$).

Chapter 4

Discussion

4.1 Summary and outlook

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 information content for medical decision-making. They are in consonance with ideas formulated previously by various authors and are compiled for the first time in the proposed list. Table 4.1 summarizes our assessment for each CBPTI plot based on our guiding 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 and biostatistics, the other focused on the present practice documented in medical journals. We found that newly developed methodological and biostatical approaches to CBPTI plots are an attempt to answer clinical questions relevant to medical decision-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 lack of presentation of statistical uncertainty, the outcome measure being given as relative unit instead of absolute unit, incorrect benchmarking, and being non-informative for medical decision-making. We encourage researchers to follow the guidelines specified in our study for improved presentation of graphics in future trial reports.

In 2014, European Medicines Agency (EMA) proposed the new guidelines focused on the investigation of subgroups in clinical research [1]. The Section 4.3 addressed the basic considerations for evaluation of heterogeneous treatment effect and associated data presentations. However, the principles ignore some key issues relevant to interpretation of results and medical decision making in data presentation. Their guidelines need to be further discussed and improved. First, it is recommended by EMA to show statistical uncertainty

by confidence intervals. The confidence intervals for binary or categorical biomarkers are straightforward. While in a complicated setting such as continuous biomarker, the approaches of showing uncertainty of treatment effect over the range of the biomarker, either pointwise confidence intervals or simultaneous confidence bands, depend on the purpose of study and should addressed at the pre-specified statistical plan. If the authors apply inappropriately, it may lead to erroneous medical decision since simultaneous confidence bands are wider than pointwise confidence intervals.

Second, 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 critical important. EMA should clearly indicate 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 the heterogeneity of treatment effect among subgroups. Incorrect benchmarking will affect the readers to make the faulty medical decision. The same idea also fits to graphical presentation in treatment-continuous biomarker interaction.

Third, both ICH E9 and the new guidelines on the investigation of subgroups proposed by EMA indicate heterogeneity of treatment effect should detect first through the addition of interaction terms to the regression models. The new guidelines further point out the sole reporting of interaction term is inadequate. It is recommended to show differences in treatment effects among subgroups. However, the guidelines for investigation of heterogeneous treatment effect are still insufficient since Huang et al [16] have demonstrated the two scenarios with the same regression coefficient of interaction term but very different biomarker performance because it depends on the scale and functional form of biomarker. Interaction plot or contrast plot fail to lead medical decision making. Huang [16] proposed a ROC curve to overcome this limitation snice 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 potential outcomes framework with severe restrictions. They assume the experimental treatment will 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 randomization assumption, we assume biomarker distributions among treatment groups are equal. Thus, treatment difference given biomarker value 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 neither in the biostatistical/methodological journals nor in the medical journals. Instead a hand search was performed. 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, the reference [20] 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 of how these plots actually support practical clinical decision-making (a question of internal/external validity). In spite of the limitations, our systematic review provides not only comprehensive methodologies on assessing CBPTI plots but also critical guiding 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 vignette called cbpti, which implements interaction plots, contrast plots, proportion of unfavorable treatment effect plots, ROC curve, and prediction curve. The main limitation of our R vignette for nonlinear function of biomarker is only employed in fractional polynomial function instead of spline function.

4.2 Conclusions

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. Although interaction plots and contrast plots are commonly used in medical literature, we would encourage researchers to employ new methods such as the selection impact curve, the marker-by-treatment predictiveness curve, risk curve, proportion of unfavorable treatment effect plot, ROC curve, and prediction curve, as such approaches answer key clinical questions relevant to medical decision-making. The proposed guiding principles in our report would be helpful for the improved presentation of CBPTI plots in future practice.

Table 4.1: The summary checklist for assessing CBPTI plots based on the guiding principles.

	Fig.1.1	Fig.3.1	Fig.3.2	Fig.3.3	Fig.3.4	Fig.3.5	Fig.3.6	Fig.3.7	Fig.3.8	Fig.3.9	Fig.3.10	Fig.3.11
Following the principles of visual display	No	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Quantifying statistical uncertainty	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No
Using absolute unit for outcome measures	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Displaying correct benchmark line	No	–	–	No	No	No	No	Yes	Yes	–	–	No
Informative content for medical decision-making	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes

Bibliography

[1] European Medicines Agency. Guidelines on the investigation of subgroups in confirmatory clinical trials. pages 1–20, 2014.

[2] Douglas G Altman, Berthold Lausen, Willi Sauerbrei, and Martin Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86(11):829–835, 1994.

[3] Douglas G Altman and Patrick Royston. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080, 2006.

[4] Rebecca Armstrong, Nicki Jackson, Jodie Doyle, Elizabeth Waters, and Faline Howes. It's in your hands: the value of handsearching in conducting systematic reviews of public health interventions. *Journal of Public Health*, 27(4):388–391, 2005.

[5] Donna L Berry, Fangxin Hong, Barbara Halpenny, Ann H Partridge, Jesse R Fann, Seth Wolpin, William B Lober, Nigel E Bush, Upendra Parvathaneni, Anthony L Back, et al. Electronic self-report assessment for cancer and self-care support: results of a multicenter randomized trial. *Journal of Clinical Oncology*, 32(3):199–205, 2014.

[6] Harald Binder, Willi Sauerbrei, and Patrick Royston. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine*, 32(13):2262–2277, 2013.

[7] J Martin Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.

[8] Marco Bonetti and Richard D Gelber. A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. *Statistics in Medicine*, 19(19):2595–2609, 2000.

[9] Tianxi Cai, Lu Tian, Peggy H Wong, and LJ Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2011.

[10] Jack Cuzick. Forest plots and the interpretation of subgroups. *The Lancet*, 365(9467):1308, 2005.

[11] Mark D Eisner. The challenge of subgroup analyses. *New England Journal of Medicine*, 355(2):211–211, 2006.

[12] Gary L Gadbury and Hari K Iyer. Unit–treatment interaction and its practical consequences. *Biometrics*, 56(3):882–885, 2000.

[13] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[14] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[15] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[16] Ying Huang, Peter B Gilbert, and Holly Janes. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, 68(3):687–696, 2012.

[17] Holly Janes, Marshall D Brown, Ying Huang, and Margaret S Pepe. An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics*, 10(1):99–121, 2014.

[18] Holly Janes, Margaret S Pepe, Patrick M Bossuyt, and William E Barlow. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*, 154(4):253–259, 2011.

[19] Rüdiger P Laubender and Ulrich Mansmann. Estimating individual treatment effects from responses and a predictive biomarker in a parallel group RCT. 2014.

[20] Shigeyuki Matsui, Richard Simon, Pingping Qu, John D Shaughnessy, Bart Barlogie, and John Crowley. Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clinical Cancer Research*, 18(21):6065–6073, 2012.

[21] Ally Olotu, Gregory Fegan, Juliana Wambua, George Nyangweso, Ken O Awuondo, Amanda Leach, Marc Lievens, Didier Leboulleux, Patricia Njuguna, Norbert Peshu, et al. Four-year efficacy of RTS, S/AS01E and its interaction with malaria exposure. *New England Journal of Medicine*, 368(12):1111–1120, 2013.

[22] Hubert Piessevaux, Marc Buyse, Michael Schlichting, Eric Van Cutsem, Carsten Bokemeyer, Steffen Heeger, and Sabine Tejpar. Use of early tumor shrinkage to predict long-term outcome in metastatic colorectal cancer treated with cetuximab. *Journal of Clinical Oncology*, 31(30):3764–3775, 2013.

[23] Stuart J Pocock, Thomas G Travison, and Lisa M Wruck. Figures in clinical trial reports: current practice & scope for improvement. *Trials*, 8(1):1–17, 2007.

[24] Stuart J Pocock, Thomas G Travison, and Lisa M Wruck. How to interpret figures in reports of clinical trials. *BMJ*, 336(7654):1166–9, 2008.

[25] Peter M Rothwell, Ziyah Mehta, Sally C Howard, Sergei A Gutnikov, and Charles P Warlow. From subgroups to individuals: general principles and the example of carotid endarterectomy. *The Lancet*, 365(9455):256–265, 2005.

[26] Patrick Royston, Mahesh KB Parmar, and Douglas G Altman. Visualizing length of survival in time-to-event studies: a complement to kaplan–meier plots. *Journal of the National Cancer Institute*, 100(2):92–97, 2008.

[27] Patrick Royston and Willi Sauerbrei. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine*, 23(16):2509–2525, 2004.

[28] Patrick Royston and Willi Sauerbrei. Interactions between treatment and continuous covariates: a step toward individualizing therapy. *Journal of Clinical Oncology*, 26(9):1397–1399, 2008.

[29] Patrick Royston and Willi Sauerbrei. Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Statistics in Medicine*, 32(22):3788–3803, 2013.

[30] Patrick Royston and Willi Sauerbrei. Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. *Statistics in Medicine*, 33(27):4695–4708, 2014.

[31] Patrick Royston, Willi Sauerbrei, and A Ritchie. Is treatment with interferon-α effective in all patients with metastatic renal carcinoma? a new approach to the investigation of interactions. *British Journal of Cancer*, 90(4):794–799, 2004.

[32] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

[33] Xiao Song and Margaret Sullivan Pepe. Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4):874–883, 2004.

[34] Spotswood L Spruance, Julia E Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–2792, 2004.

[35] Xin Sun, Matthias Briel, Jason W Busse, John J You, Elie A Akl, Filip Mejza, Malgorzata M Bala, Dirk Bassler, Dominik Mertz, Natalia Diaz-Granados, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*, 344:e1553, 2012.

[36] Xin Sun, Matthias Briel, Stephen D Walter, and Gordon H Guyatt. Is a subgroup effect believable? updating criteria to evaluate the credibility of subgroup analyses. *BMJ*, 340:c117, 2010.

[37] Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[38] Eric Van Cutsem, Claus-Henning Köhne, Erika Hitre, Jerzy Zaluski, Chung-Rong Chang Chien, Anatoly Makhson, Geert D'Haens, Tamás Pintér, Robert Lim, György Bodoky, et al. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *New England Journal of Medicine*, 360(14):1408–1417, 2009.

[39] Giuseppe Viale, Anita Giobbie-Hurder, Meredith M Regan, Alan S Coates, Mauro G Mastropasqua, Patrizia Dell'Orto, Eugenio Maiorano, Gaëtan MacGrogan, Stephen G Braye, Christian Öhlschlegel, 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. *Journal of Clinical Oncology*, 26(34):5569–5575, 2008.

[40] Antonio C Wolff, Ann A Lazar, Igor Bondarenko, August M Garin, Stephen Brincat, Louis Chow, Yan Sun, Zora Neskovic-Konstantinovic, Rodrigo C Guimaraes, Pierre Fumoleau, 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. *Journal of Clinical Oncology*, 31(2):195–202, 2013.

[41] Greg Yothers, Michael J O'Connell, Mark Lee, Margarita Lopatin, Kim M Clark-Langone, Carl Millward, Soonmyung Paik, Saima Sharif, Steven Shak, and Norman Wolmark. 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. *Journal of Clinical Oncology*, 31(36):4512–4519, 2013.

Acknowledgements

Figures 3.1, 3.2, and 3.3 are reproduced with permission from the Journal of Clinical Oncology and originally published by the American Journal of Clinical Oncology. [Figure 3.1: 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 3.2: 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 3.3: 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 1.1 and 3.5 are reproduced with permission from Statistics in Medicine by John Wiley & Sons, Ltd. [Figure 1.1: Bonetti M, Gelber RD: Statistics in Medicine Vol. 19 (Issue 19), date: October 15, 2000] [Figure 3.5: Royston P, Sauerbrei W: Statistics in Medicine Vol. 23 (Issue 16), date: August 30, 2004]

Figure 3.4 is reproduced with permission from Biostatistics by Oxford University Press. [Figure 3.4: Cai T, Tian L, Wong PH, Wei LJ: Biostatistics Vol. 12 (Issue 2), date: September 28, 2010]

Figure 3.6 is reproduced with permission from the New England Journal of Medicine by Massachusetts Medical Society. [Figure 3.6: 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 3.7, the dataset was provided by Prof. Willi Sauerbrei.

Figures 3.8 and 3.11 are reproduced with permission from Biometrics by John Wiley & Sons, Inc. [Figure 3.8: Song X, Pepe MS: Biometrics Vol. 60 (Issue 4), date: December of 2004]. [Figure 3.11: Huang Y, Gilbert PB, Janes H: Biometrics Vol. 68 (Issue 3), date: September of 2012].

Figure 3.9 is reproduced with permission from the Annals of Internal Medicine by the American College of Physicians. [Figure 3.9: Janes H, Pepe MS, Bossuyt PM, Barlow

WE: Annals of Internal Medicine Vol. 154 (Issue 4), date: February 15, 2011]

Figure 3.10 is reproduced with permission from The International Journal of Biostatistics by De Gruyter. [Figure 3.10: Janes H, Brown MD, Huang Y, Pepe MS: The International Journal of Biostatistics Vol. 10 (Issue 1), date: April 2, 2014]

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dean's Office
Medical Faculty

Affidavit

Shen, Yu-Ming

Surname, first name

Otkerstraße 7d

Street

81547, München

Zip code, town

Germany

Country

I hereby declare, that the submitted thesis entitled

Graphical presentation of patient-treatment interaction elucidated by continuous biomarker

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

Place, date

Signature doctoral candidate