Contemporary Clinical Trials Communications 5 (2017) 107–115

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

# Innovative methods for the identification of predictive biomarker signatures in oncology: Application to bevacizumab

Paul Delmar [a, *], Cornelia Irl [b], Lu Tian [c]

[a] Department of Biostatistics, F. Hoffmann-La Roche Ltd., Basel, Switzerland
[b] Department of Biostatistics, Genentech Inc., South San Francisco, CA, USA
[c] Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA, USA

ARTICLE INFO

*Article history:*
Received 18 December 2015
Received in revised form
6 December 2016
Accepted 17 January 2017
Available online 19 January 2017

*Keywords:*
Bevacizumab
Breast cancer
Personalized medicine
Subgroup analysis
Multivariate statistical classification
Machine learning

ABSTRACT

Current methods for subgroup analyses of data collected from randomized clinical trials (RCTs) may lead to false-positives from multiple testing, lack power to detect moderate but clinically meaningful differences, or be too simplistic in characterizing patients who may benefit from treatment. Herein, we present a general procedure based on a set of newly developed statistical methods for the identification and evaluation of complex multivariate predictors of treatment effect. Furthermore, we implemented this procedure to identify a subgroup of patients who may receive the largest benefit from bevacizumab treatment using a panel of 10 biomarkers measured at baseline in patients enrolled on two RCTs investigating bevacizumab in metastatic breast cancer. Data were collected from patients with human epidermal growth factor receptor 2 (HER2)-negative (AVADO) and HER2-positive (AVEREL) metastatic breast cancer. We first developed a classification rule based on an estimated individual scoring system, using data from the AVADO study only. The classification rule takes into consideration a panel of biomarkers, including vascular endothelial growth factor (VEGF)-A. We then classified the patients in the independent AVEREL study into patient groups according to "promising" or "not-promising" treatment benefit based on this rule and conducted a statistical analysis within these subgroups to compute point estimates, confidence intervals, and p-values for treatment effect and its interaction. In the group with promising treatment benefit in the AVEREL study, the estimated hazard ratio of bevacizumab versus placebo for progression-free survival was 0.687 (95% confidence interval [CI]: 0.462–1.024, $p = 0.065$), while in the not-promising group the hazard ratio (HR) was 1.152 (95% CI: 0.526–2.524, $p = 0.723$). Using the median level of VEGF-A from the AVEREL study to divide the study population, then the HR becomes 0.711 (95% CI: 0.435–1.163, $p = 0.174$) in the promising group and 0.828 (95% CI: 0.496–1.380, $p = 0.468$) in the not-promising group. Similar results were obtained with the median VEGF-A levels from the AVADO study ("promising" group: HR = 0.709, 95%CI: 0.444–1.133, $p = 0.151$; "not-promising" group: HR = 0.851, 95% CI: 0.497–1.458, $p = 0.556$). Our analysis shows it is feasible to employ statistical methods for empirically constructing and validating a scoring system based on a panel of biomarkers. This scoring system can be used to estimate the treatment effect for individual patients and identify a subgroup of patients who may benefit from treatment. The proposed procedure can provide a general framework to organize many statistical methods (existing or to be developed) into a coherent set of analyses for the development of personalized medicines and has the potential of broad applications.



## 1. Introduction

Randomized clinical trials are designed to assess the efficacy of a new treatment compared with placebo or standard of care. Oftentimes, in addition to the main comparison of the overall population enrolled in the study, subgroup analyses are performed to examine whether the benefit of the new treatment is consistent across patient populations [1]. Specifically, subgroup analyses aim

*Abbreviations:* FGF, fibroblast growth factor; FLT, fms-like tyrosine kinase; HER2, human epidermal growth factor receptor 2; MRST, in mean restricted survival time; RCT, randomized clinical trial; VEGF, vascular endothelial growth factor.

* Corresponding author. Department of Biostatistics, F.Hoffmann-La Roche Ltd., 4070 Basel, Switzerland.
*E-mail addresses:* paul.delmar@roche.com (P. Delmar), cornelia.irl@roche.com (C. Irl), lutian@stanford.edu (L. Tian).

http://dx.doi.org/10.1016/j.conctc.2017.01.007

to estimate and test the treatment effect on pre-determined subgroups. The subgroups are usually characterized by simple criterions measured at baseline, such as sex, race, comorbidities, and pre-existing treatment status. The final results are often presented graphically in a forest plot (e.g., Fig. 1), where each tree represents the point, as well as the interval estimates of the treatment effect within a subgroup. If one or several trees stand(s) out of the forest, this may indicate non-homogeneity of the treatment effect.

This simplicity, however, may be misleading [2,3]. The first difficulty associated with subgroup analyses is multiple testing [4]. If one tries to estimate the treatment effect in a sufficiently large number of subgroups, there will always be significant findings. This opens the door for subjective interpretation of the subgroups identified based on the significance level or the point estimator itself: it could be either a simple false-positive result due to multiple testing or a promising subgroup worthy of further investigation. Various statistical adjustments have been proposed but are rarely used in practice for good reasons [5]. For example, the Bonferroni correction is one of the most robust approaches to ensure that the treatment effect in at least one of the identified subgroups truly exists with the claimed significance level [6]. However, the adjustment is highly conservative and may fail to detect a moderate subgroup-specific treatment effect. This raises the second difficulty in subgroup analyses, i.e., lack of power to detect moderate yet clinically meaningful treatment effects [3]. Finally, the definition of the pre-defined subgroup may be too simplistic to characterize patients who may (or may not) benefit from the treatment. If we are willing to consider subgroups defined by a combination of characteristics, the number of candidate subgroups increases very rapidly, exacerbating the difficulties associated with multiple testing and lack of statistical power. For example, 10 binary characteristics can define up to 2048 different subgroups of patients. Even after acknowledging that some subgroups may be too small to be of interest, it is likely that we still need to deal with hundreds of subgroups. When some of the characteristics are continuous, such as systolic blood pressure or gene expression level, there are an infinite number of subgroups and it becomes infeasible to conduct subgroup analyses. More sophisticated methods that allow automatic identification of the subgroups of interest are needed [7–9].

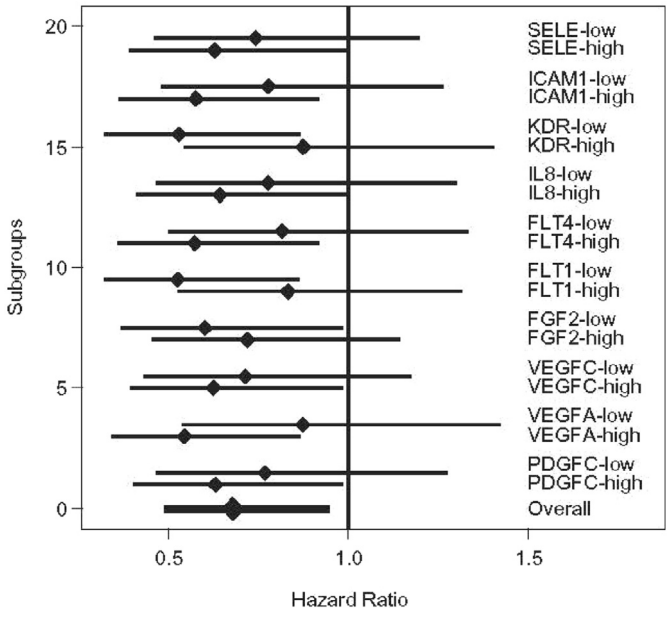


**Fig. 1.** Forest plot for subgroup analysis in AVADO study. The high and low groups are defined using the median of the corresponding biomarkers.

In light of these drawbacks of the simple subgroup analyses, there are many recent developments in statistical methodology for personalized medicine [7–19]. Among them, many adopt various modern machine learning techniques to relax conventional statistical model assumptions [11,14–19].

However, most of these recent developments are fragmentary and there is no practical guideline for conducting the complete statistical analysis for personalized medicine. For example, in the presence of multiple approaches for estimating personalized treatment effect and even different metrics for quantifying the personalized treatment effect, there is a lack of methods for selecting the optimal approach.

We have identified three goals for statistical methods in personalized medicine: (1) estimating the treatment effect for the individual patient, i.e., the individualized treatment effect [9,10,14–19], (2) building a classification rule for identifying patients who may (or may not) benefit from the treatment, or (3) making valid statistical inferences about treatment effect in the identified subgroup.

In this paper, we propose a coherent stage-wise procedure for addressing all three objectives. It has a clearly defined target at each step. The procedure is also flexible and can easily be extended to leverage new or future developments in the field. This procedure will be illustrated by analyzing the data from two randomized clinical oncology trials conducted by Hoffmann-La Roche Inc. In both trials, the overall comparisons showed moderate treatment effect in the entire study population and it is desirable to identify a subgroup of patients having more substantial treatment benefit [26,27]. However, the simple subgroup analysis failed to detect and confirm the existence of the heterogeneous treatment effect [28].

## 2. Methods

### 2.1. Procedure for subgroup selection

The procedure consists of two major steps: training and testing. The outcome of the training step is a classification rule for selecting a subgroup of patients based on baseline features including biomarker levels, demographic information, comorbidities, etc. The classification rule can be complex and depends on multiple features. The outcome of the testing step is the verification and evaluation of the treatment effect in the subgroup identified by the classification rule, as well as in the complementary subgroup. In the ideal case, there are data from two randomized clinical trials and we use the first for training (Part I) and the second for testing (Part II). If all the data are from a single trial, we need to split the data into two non-overlapping parts (Parts I and II).

### 2.2. Training step

In this stage, we estimate the treatment effect for individual patients and construct a classification rule for selecting patients with promising treatment effect. However, several estimation methods can be used and we need to select the optimal one based on the data. To this end, the estimation and validation steps need to be built within the training step. Specifically, the training data will be randomly split into two parts: the first part (Part I-E) will be used to estimate the treatment effect for individual patients with different methods; the second part (Part I-V) will be used to evaluate the performance of each of the estimated treatment effects in stratifying patient population into strata of different treatment effects.