Do Surrogate Endpoints Better Correlate with Overall Survival in Studies That Did Not Allow for Crossover or Reported Balanced Postprogression Treatments?
Hashim, Mahmoud; Pfeiffer, Boris M; Bartsch, Robert; Postma, Maarten; Heeg, Bart

Published in:
Value in Health

DOI:

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Mahmoud Hashim, MPH1,*, Boris M. Pfeiffer, PhD2, Robert Bartsch, MSc1, Maarten Postma, PhD3, Bart Heeg, PhD1

1Ingress-health, Rotterdam, The Netherlands; 2Merck KgaA, Darmstadt, Germany; 3University of Groningen, Groningen, The Netherlands

ABSTRACT

Background: In previous studies, correlation between overall survival (OS) and surrogate endpoints like objective response rate (ORR) or progression-free survival (PFS) in advanced non-small cell lung cancer (NSCLC) was poor. This can be biased by crossover and postprogression treatments. Objectives: To evaluate the relationship between these two surrogate endpoints and OS in advanced NSCLC studies that did not allow for crossover or reported balanced post-progression treatments. Methods: A systematic review in patients with advanced NSCLC receiving second- and further-line therapy was performed. The correlation between the absolute difference in ORR or median PFS (mPFS) and the absolute difference in median OS (mOS) was assessed using the correlation coefficient (R) and weighted regression models. The analysis was repeated in predefined data cuts based on crossover and balance of postprogression treatments. When the upper limit of R's 95% confidence interval (CI) was more than 0.7, the surrogate threshold effect (STE) was estimated. Results: In total, 146 randomized clinical trials (43,061 patients) were included. The mean ORR, mPFS, and mOS were 12.2% ± 11.2%, 3.2 ± 3.1 months, and 9.6 ± 4.1 months, respectively. The correlation coefficients of ORR and mPFS were 0.181 (95% CI 0.016–0.337) and 0.254 (95% CI 0.074–0.418), respectively, with mOS. Nevertheless, in trials that did not allow crossover and reported balanced postprogression treatments, the correlation coefficients of ORR and mPFS were 0.528 (95% CI 0.081–0.798) and 0.778 (95% CI 0.475–0.916), respectively, with mOS. On the basis of STE estimation, in trials showing significant treatment effect size of 41.0% or more ORR or 4.15 or more mPFS months, OS benefit can be expected with sufficient certainty. Conclusions: Crossover and postprogression treatments may bias the relationship between surrogate endpoints and OS. Presented STE calculation can be used to interpret treatment effect on either ORR or PFS when used as primary endpoints.

Keywords: crossover, non–small cell lung cancer, overall survival, surrogate endpoints validation.

Introduction

Overall survival (OS) is the criterion standard endpoint in cancer trials and is used to establish clinical benefit in support of regulatory and reimbursement applications [1–4]. Nevertheless, trials using OS as a primary endpoint need substantial sample sizes and extensive follow-up. In addition, the effects of crossover or unbalanced postprogression treatments may introduce bias or underestimate the treatment effect on OS [5,6]. An alternative surrogate endpoint for OS is progression-free survival (PFS). Regulatory agencies endorse PFS as a relevant endpoint in cancer trials [1,2,7]. In contrast to OS, PFS is not sensitive to postprogression treatments and has the advantage of assessing the duration of tumor response [5]. Objective response rate (ORR) is another potential surrogate endpoint. Compared with PFS, ORR does not assess response duration. The use of PFS and ORR as surrogate endpoints for OS would require that they be validated for this use [8]. Nevertheless, uncertainties regarding their association with OS and the potential for bias due to subjectivity in the assessment of ORR and PFS limit their use [7].

To our knowledge, only the Institute for Quality and Efficiency in Health Care (IQWiG) has issued a guidance document for surrogate endpoint validation in oncology [4]. The IQWiG recommends a stringent definition of surrogacy on the basis of the correlation coefficient (R). IQWiG states that if the lower limit of the 95% confidence interval (CI) of R is 0.85 or higher, validity of

E-mail: mahmoud.hashim@gress-health.com.
the surrogate is suggested, but that the surrogate is not valid if the upper limit of the 95% CI is 0.7 or less [4]. Otherwise, the validity of the surrogate remains unclear; in this situation, IQWiG recommends estimating the surrogate threshold effect (STE) [4,9]. STE is defined as the minimum treatment effect on the surrogate necessary to predict a statistically significant nonzero effect on the true endpoint [9]. STE can be used to interpret the treatment effect on the surrogate endpoint.

A few studies in non–small cell lung cancer (NSCLC) have investigated the surrogacy of ORR or PFS to OS at the trial level [6,10–13]. These studies reported low correlations between PFS or ORR and OS. None of them included a stratified analysis based on the exclusion of studies allowing crossover or reporting unbalanced postprogression treatments. Stratifying studies on the basis of crossover has been done in other tumor types [14–16]. Delea et al. [15] assessed the surrogacy of PFS to OS in metastatic renal cell carcinoma trials. The correlation coefficient was greater in studies that did not allow/require crossover versus those that did allow/require crossover: correlation coefficients were estimated to be 0.50 and 0.28, respectively. Similarly, and to a less extent, greater correlation coefficients were observed in endpoint validation studies for metastatic melanoma and metastatic colorectal cancer after the removal of studies that did not allow/require crossover [14,16]. Hence, investigating the effect of crossover and postprogression treatments on the surrogacy of ORR or PFS to OS in NSCLC is warranted.

This study aimed to evaluate ORR and PFS as surrogate endpoints for OS in trials involving patients with advanced NSCLC receiving second- and further-line therapy. Then, the impact of crossover and unbalanced postprogression treatments on surrogacy was assessed.

Methods

Systematic Literature Review

The systematic literature review was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement [17]. Two different bibliographic databases, PubMed and Embase, were used to identify published randomized clinical trials involving patients with stage IIIIB/IV NSCLC receiving second- and further-line therapy. The search was conducted on July 28, 2016; no limitation on publication date was imposed.

A detailed search strategy (search syntax and eligibility criteria) is presented in Appendix Table 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011. One investigator reviewed the titles/abstracts of retrieved articles sequentially using the predefined eligibility criteria (see Appendix Table 1 in Supplemental Materials). Subsequently, two investigators reviewed the full text of any article that appeared to meet the eligibility criteria; disagreement was resolved by consulting with a third investigator. References in publications reviewed at the full-text stage were evaluated to identify further relevant trials.

Upon agreement on the final list of included trials, one investigator extracted data from the included trials into a predefined Microsoft Excel template. Subsequently, another investigator validated the extracted data by re-extracting them. The following data were extracted: trial identification items (e.g., PubMed identifier, first author, year, trial phase, registration identifier, and trial acronym), interventions and target population, basic patient and disease characteristics (e.g., age, sex, performance status, disease stage, histology, metastasis, and number of previous lines of therapy), additional information (e.g., use of biomarkers and crossover), and data needed for endpoint validation (number of patients in each treatment arm, ORR, PFS, and OS). Risk of bias in individual studies was assessed using the Jadad scale [18].

Assessment of Publication Bias

The risk of bias across studies was assessed using funnel plots. In this study, trial size as a measure of precision was plotted on the y-axis, and treatment effect (absolute difference) on ORR, PFS, and OS was plotted on the x-axis. In the absence of publication bias, the plot should resemble a symmetrical inverted funnel [19].

Statistical Analysis

Primary analysis

The relationship between the absolute difference in ORR and median PFS (mPFS) and the absolute difference in median OS (mOS) was assessed using the correlation coefficient (R) and weighted linear regression models. A weighted linear regression model was fitted for the following two analyses: treatment effect on ORR, with the absolute difference in ORR (%) as an independent variable (predictor) and the treatment effect on OS (absolute difference in mOS in months) as a dependent variable; and treatment effect on PFS, with the absolute difference in mPFS (months) as an independent variable (predictor) and the treatment effect on OS (absolute difference in mOS in months) as a dependent variable. Analyses were weighted by trial size, as in previous endpoint validation studies [10,13,15,20–22].

Analyses were repeated using the absolute difference in ORR (%) or PFS hazard ratio (HR) and OS-HR because HRs might capture treatment effects not captured by median survival times. We carried out log transformation of HR. Log transformation can be used to make right-skewed distributions less skewed. Treatment effect on ORR is usually reported as the absolute difference in ORR (%). For that reason and for the ease of interpretation, we used it in both analyses with OS (mOS and OS-HR). Residual versus predicted plots were inspected and diagnostic tests for normality and heteroscedasticity (nonconstant error variance) were carried out to assess consistency with the assumptions of linear regression.

First, the analysis was conducted for all trials. Trials that had allowed crossover or in which postprogression treatments were unbalanced could underestimate OS benefit and subsequently bias surrogacy evaluation. Typically, phase III trials are adequately powered for endpoints such as PFS and OS, whereas phase II trials tend to be smaller and powered for safety endpoints or ORR. Thus, phase III trials might provide more information regarding the treatment effect on these endpoints. Therefore, second, on the basis of reported postprogression treatments, we examined trial-level surrogacy in all phase III trials (data cut A), in phase III trials excluding those with per-protocol crossover (data cut B), in phase III trials excluding those with both per-protocol and off-protocol crossover (data cut C), and in phase III trials excluding those with crossover, unbalanced postprogression treatments, or no information with regard to postprogression treatments (data cut D).

Trials that reported both the independent (the surrogate endpoint) and the dependent (the true endpoint) variables in both treatment arms were included in the analyses. For trials that included more than two treatment arms, the experimental arm was compared with a randomly chosen control arm within the same study to avoid analysis of correlated data, that is, including a treatment arm twice in the analysis. For trials that reported response in the evaluable population rather than in the intention-to-treat population, the denominator was adjusted to indicate the intention-to-treat population.

Assessing surrogacy and STE estimation

In cases in which the validity of the surrogate endpoint is deemed to be “unclear” following IQWiG guidelines [4], STE estimation is recommended to interpret treatment effect on the
Table 1 – Basic population characteristics in all included trials and prespecified data cuts based on reported postprogression therapies.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All trials (n = 146)</th>
<th>Data cut A (n = 59)</th>
<th>Data cut B (n = 54)</th>
<th>Data cut C (n = 38)</th>
<th>Data cut D (n = 18)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid (n)</td>
<td>Mean ± SD</td>
<td>Valid (n)</td>
<td>Mean ± SD</td>
<td>Valid (n)</td>
</tr>
<tr>
<td>Age (y), median*</td>
<td>288</td>
<td>61.3 ± 3.2</td>
<td>116</td>
<td>61.2 ± 3.3</td>
<td>106</td>
</tr>
<tr>
<td>Male (%)</td>
<td>292</td>
<td>63.7 ± 13.4</td>
<td>118</td>
<td>64.9 ± 13.5</td>
<td>108</td>
</tr>
<tr>
<td>ECOG 0 or 1 (%)</td>
<td>278</td>
<td>90.2 ± 12.2</td>
<td>116</td>
<td>91.2 ± 8</td>
<td>106</td>
</tr>
<tr>
<td>Adenocarcinoma (%)</td>
<td>238</td>
<td>63.3 ± 18.6</td>
<td>96</td>
<td>65.2 ± 18</td>
<td>90</td>
</tr>
<tr>
<td>ORR (%)</td>
<td>290</td>
<td>12.2 ± 11.2</td>
<td>118</td>
<td>12.4 ± 12.7</td>
<td>108</td>
</tr>
<tr>
<td>PFS (mo)</td>
<td>236</td>
<td>3.2 ± 1.3</td>
<td>100</td>
<td>3.2 ± 1.4</td>
<td>90</td>
</tr>
<tr>
<td>OS (mo)</td>
<td>282</td>
<td>9.6 ± 4.1</td>
<td>118</td>
<td>10 ± 4.2</td>
<td>108</td>
</tr>
<tr>
<td>Jadad scale†</td>
<td>146</td>
<td>2.7 ± 1.0</td>
<td>59</td>
<td>3 ± 1.1</td>
<td>54</td>
</tr>
<tr>
<td>Sample size†</td>
<td>146</td>
<td>294.9 ± 321.2</td>
<td>59</td>
<td>548.3 ± 382.3</td>
<td>54</td>
</tr>
</tbody>
</table>

Note. Data cuts based on reported postprogression treatments: data cut A, phase III trials; data cut B, phase III trials excluding those with per-protocol crossover; data cut C, phase III trials excluding those with both per-protocol and off-protocol crossover; and data cut D, phase III trials excluding those with crossover, unbalanced postprogression treatments or insufficient information.

ECOG, Eastern Cooperative Oncology Group, performance status; n, number of observations; ORR, objective response rate; OS, overall survival; PFS, progression-free survival.

* Valid number of treatment arms with reported observation.
† Valid number of trials with reported observation.
surrogate endpoint. STE is the minimum treatment effect on the surrogate necessary to predict a statistically significant nonzero effect on the true endpoint [9]. The STE calculation allows threshold values for the decision as to whether an observed effect on the surrogate would predict (with sufficient certainty) an effect on the endpoint of interest to be specified [9]. To draw such a conclusion, the lower confidence limit of the treatment effect on the surrogate must be larger than the STE. To calculate the STE, the regression line was plotted using the weighted linear regression equation. Then, 95% prediction intervals were plotted. The value on the x-axis, the treatment effect on the surrogate, at which the lower limit of the prediction interval (upper limit in the case of relative treatment effect) meets a point corresponding to 0 on the y-axis (zero effect on the true endpoint) is the STE [9]. With stronger correlation between the surrogate endpoint and the hard endpoint, it is easier to reach the STE, for example, lower incremental mPFS months or closer to 1 PFS-HR.

Additional analyses
In addition to the likely bias due to the existence of crossover and/or unbalanced postprogression treatments, other trial or patient’s characteristics may bias the quantitative relationship between surrogate endpoints and OS. Thus, first we fitted a multivariate weighted linear regression model. Such analysis would investigate whether the likely bias caused by crossover and/or unbalanced postprogression treatments still holds after adjustment for other available variables. The analysis was run only for the absolute difference in mPFS in phase III trials. The initial list of candidate independent (predictor) variables, in addition to ΔPFS, included median age, male (%), Eastern Cooperative Oncology Group (performance status) 0 or 1 (%), adenocarcinoma (%), endothelial growth factor receptor tyrosine kinase inhibitor (EGFR-TKI) treatment (dummy variable), OS primary endpoint (dummy variable), publication year, assessed Jadad scale, and data cut D (dummy variable). Crossover and postprogression treatments happen after progression. Therefore, the simultaneous influence of ΔPFS and data cut D on OS is not additive. This justifies adding an interaction term between them in the regression model.

Second, a logistic regression model was fitted with data cut D (yes = 1; no = 0) as the dependent variable. The same trial and patient’s characteristics were considered as independent variables. This analysis should give more insight on the differences between studies included and excluded in the data cut D.

All analyses were carried out using the statistical software package R version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria) and using Package “Surrogate” version 0.1-67 [23].

Results
Systematic Literature Review
Of 6274 potentially relevant publications identified, 299 hits qualified for full-text screening. After the full-text screening, 146 trials (43,061 patients) fulfilled the eligibility criteria and were

### Table 2 – Association between treatment effect on ORR and PFS with OS.

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Dependent variable</th>
<th>Subgroup</th>
<th>No. of patients</th>
<th>Correlation coefficient (R)</th>
<th>95% CI</th>
<th>STE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔORR (%)</td>
<td>ΔOS, median (mo)</td>
<td>All trials</td>
<td>140</td>
<td>41,725</td>
<td>0.181</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut A</td>
<td>59</td>
<td>32,348</td>
<td>0.131</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut B</td>
<td>54</td>
<td>30,654</td>
<td>0.361</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut C</td>
<td>38</td>
<td>22,574</td>
<td>0.445</td>
<td>0.146</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut D</td>
<td>18</td>
<td>13,349</td>
<td>0.528</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All trials</td>
<td>76</td>
<td>30,570</td>
<td>0.172</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut A</td>
<td>44</td>
<td>26,549</td>
<td>0.374</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut B</td>
<td>41</td>
<td>25,534</td>
<td>0.399</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut C</td>
<td>27</td>
<td>18,854</td>
<td>0.521</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut D</td>
<td>17</td>
<td>13,194</td>
<td>0.164</td>
<td>0.000</td>
</tr>
<tr>
<td>PFS, median (mo)</td>
<td>OS, median (mo)</td>
<td>All trials</td>
<td>114</td>
<td>35,729</td>
<td>0.254</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut A</td>
<td>50</td>
<td>27,579</td>
<td>0.260</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut B</td>
<td>45</td>
<td>25,885</td>
<td>0.438</td>
<td>0.166</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut C</td>
<td>30</td>
<td>18,634</td>
<td>0.741</td>
<td>0.520</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut D</td>
<td>17</td>
<td>13,194</td>
<td>0.778</td>
<td>0.475</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All trials</td>
<td>73</td>
<td>29,907</td>
<td>0.402</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut A</td>
<td>42</td>
<td>25,386</td>
<td>0.463</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut B</td>
<td>39</td>
<td>24,371</td>
<td>0.461</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut C</td>
<td>25</td>
<td>17,691</td>
<td>0.694</td>
<td>0.412</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data cut D</td>
<td>17</td>
<td>13,194</td>
<td>0.698</td>
<td>0.326</td>
</tr>
</tbody>
</table>

Note. Data cuts based on reported postprogression treatments: data cut A, all phase III trials regardless of postprogression treatments; data cut B, phase III trials excluding those with per-protocol crossover; data cut C, phase III trials excluding those with both per-protocol and off-protocol crossover; and data cut D, phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information. Δ, absolute difference; CI, confidence interval; HR, hazards ratio; NA, not available; ORR, objective response rate; OS, overall survival; PFS, progression-free survival; STE, surrogate threshold effect.

† ΔORR (%).
‡ ΔPFS median months.
§ PFS-HR.
Fig. 1 – Relationship between ΔORR (x-axis) and ΔOS (y-axis): (A) primary analysis (all trials); (B) phase III trials (data cut A); (C) phase III trials excluding those with per-protocol crossover (data cut B); (D) phase III trials excluding those with both per-protocol and off-protocol crossover (data cut C); and (E) phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information (data cut D). The solid line is the regression line. Red dashed lines are the upper and lower limits of the 95% confidence interval. Black dashed lines are the upper and lower bands of the 95% prediction intervals. Circle size is proportionate to trial size. ORR, objective response rate; OS, overall survival.
Fig. 2 – Relationship between ΔPFS (x-axis) and ΔOS (y-axis): (A) primary analysis (all trials); (B) phase III trials (data cut A); (C) phase III trials excluding those with per-protocol crossover (data cut B); (D) phase III trials excluding those with both per-protocol and off-protocol crossover (data cut C); and (E) phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information (data cut D). The solid line is the regression line. Red dashed lines are the upper and lower limits of the 95% confidence interval. Black dashed lines are the upper and lower limits of the 95% prediction intervals. Circle size is proportionate to trial size. OS, overall survival; PFS, progression-free survival.
included in the primary analysis (see Appendix Figure 1 and Appendix Tables 2 and 3 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011). Table 1 presents basic population characteristics in all included trials and in data cuts defined on the basis of postprogression treatments. Among all phase III trials (n = 59), 5 and 16 studies reported per-protocol and off-protocol crossover, respectively. These studies were excluded from data cuts A and B, respectively. Seven trials reported unbalanced postprogression treatments and 13 trials failed to report any information with regard to crossover or postpro-
gression treatments. These studies were later excluded from data cut C.

In all treatment arms, combination therapy was the most frequent intervention (74 treatment arms) and docetaxel was the most frequent monotherapy intervention (57 treatment arms), followed by EGFR-TKIs (erlotinib/gefitinib/afatinib, 45 treatment arms) and pemetrexed (25 treatment arms). Thirty-four trials recruited exclusively Asian patients (2874 patients). A total of 87 trials involving 10,713 patients were phase III trials, whereas 59 trials involving 32,348 patients were phase II trials. The Jadad scale for individual trials was generally low (see Table 2 in Supplemental Materials); this is because most of the trials were open-label trials and sufficient information about randomization methods was not reported.

A visual examination of the funnel plots (see Appendix Figures 2–4 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011) shows that the risk of publication bias can be unlikely.

### Primary Analysis

#### Analysis of ORR as a surrogate for OS

One hundred forty trials (41,725 patients) reported both ORR and mOS in both treatment arms. The correlation coefficient between ΔORR and ΔOS was 0.181 (95% CI 0.016–0.337) (Table 2; Fig. 1). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, C, and D), the association between ΔORR and ΔOS becomes stronger: data cut A, \( R = 0.131 \) (95% CI 0.000–0.375); data cut B, \( R = 0.361 \) (95% CI 0.103–0.573); data cut C, \( R = 0.445 \) (95% CI 0.146–0.669); and data cut D, \( R = 0.528 \) (95% CI 0.081–0.798). In data cut D, the upper limit of \( R' \)'s 95% CI is more than 0.7; therefore, STE was estimated to be 41% in this data cut (see Fig. 1).

Seventy-six trials (30,570 patients) reported both ORR and OS-HR in both treatment arms. The correlation coefficient between ΔORR and log(OS-HR) was 0.172 (95% CI 0.000–0.383). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, and C), association between ΔORR and log(OS-HR) becomes stronger: data cut A, \( R = 0.374 \) (95% CI 0.086–0.604); data cut B, \( R = 0.399 \) (95% CI 0.104–0.629); and data cut C, \( R = 0.521 \) (95% CI 0.175–0.752). In data cut C, the upper limit of \( R' \)'s 95% CI is more than 0.7; therefore, STE was estimated to be 55% in this data cut. This association did not achieve statistical significance in data cut D (see Appendix Figure 5 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011).

#### Analysis of PFS as a surrogate for OS

One hundred fourteen trials (35,729 patients) reported both mPFS and mOS in both treatment arms. The correlation coefficient between ΔPFS and ΔOS was 0.254 (95% CI 0.074–0.418) (Table 2; Fig. 2). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, C, and D), association between ΔPFS and ΔOS becomes stronger: data cut A, \( R = 0.260 \) (95% CI 0.000–0.502); data cut B, \( R = 0.438 \) (95% CI 0.166–0.649); data cut C, \( R = 0.741 \) (95% CI 0.520–0.869); and data cut D, \( R = 0.778 \) (95% CI 0.475–0.916). In data cuts C and D, the upper limit of \( R' \)'s 95% CI is more than 0.7; therefore, STE was estimated to be 3.7 and 4.2 incremental mPFS months, respectively, in these data cuts (see Fig. 2).

Seventy-three trials (29,907 patients) reported both PFS-HR and OS-HR. The correlation coefficient between log(PFS-HR) and log(OS-HR) was 0.402 (95% CI 0.190–0.579). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, C, and D), association between log(PFS-HR) and log(OS-HR) becomes stronger: data cut A, \( R = 0.463 \) (95% CI 0.185–0.672); data cut B, \( R = 0.461 \) (95% CI 0.170–0.678); data cut C, \( R = 0.694 \) (95% CI 0.412–0.855); and data cut D, \( R = 0.698 \) (95% CI 0.326–0.882). In data cuts C and D, the upper limit of \( R' \)'s 95% CI is more than 0.7; therefore, STE was estimated to be 0.24 and 0.22

---

**Table 3 – Relationship between ΔPFS and ΔOS in phase III trials adjusted for patient and trial characteristics (weighted multivariate linear regression model).**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>P value</th>
<th>95% CI Lower limit</th>
<th>95% CI Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔPFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data cut D (yes)</td>
<td>1.129</td>
<td>0.001</td>
<td>0.531</td>
<td>1.727</td>
</tr>
<tr>
<td>Data cut D (no)</td>
<td>0.379</td>
<td>0.071</td>
<td>−0.035</td>
<td>0.794</td>
</tr>
<tr>
<td>Age (y), median</td>
<td>0.170</td>
<td>0.018</td>
<td>0.031</td>
<td>0.308</td>
</tr>
<tr>
<td>ECOG 0 or 1 (%)</td>
<td>1.236</td>
<td>0.654</td>
<td>−4.942</td>
<td>6.814</td>
</tr>
<tr>
<td>Adenocarcinoma (%)</td>
<td>−1.599</td>
<td>0.287</td>
<td>−4.612</td>
<td>1.414</td>
</tr>
<tr>
<td>EGFR treatment</td>
<td>−0.242</td>
<td>0.546</td>
<td>−1.051</td>
<td>0.567</td>
</tr>
<tr>
<td>OS primary end point</td>
<td>−0.240</td>
<td>0.570</td>
<td>−1.095</td>
<td>0.615</td>
</tr>
<tr>
<td>Publication year</td>
<td>−0.127</td>
<td>0.219</td>
<td>−0.334</td>
<td>0.079</td>
</tr>
<tr>
<td>Jadad scale</td>
<td>−0.227</td>
<td>0.273</td>
<td>−0.641</td>
<td>0.188</td>
</tr>
<tr>
<td>Constant</td>
<td>246.523</td>
<td>0.230</td>
<td>−164.511</td>
<td>657.556</td>
</tr>
</tbody>
</table>

Note. Data cut D is defined as phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information.

CI, confidence interval; ECOG, Eastern Cooperative Oncology Group, performance status; EGFR, epidermal growth factor receptor; OS, overall survival; PFS, progression-free survival.

† Interaction term between PFS and data cut D is added to the model.

‡ High correlation between “Male” and “Adenocarcinoma” was observed (\( r = −0.808 \)). On the basis of collinearity diagnostics, high variance inflation factor (\( >5 \)) and low tolerance (\( <0.2 \)) values were observed. Therefore, we omitted the “Male” variable from the regression model.
PFS-HR, respectively, in these data cuts (see Appendix Figure 6 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011).

Additional Analysis

High correlation between “Male” and “Adenocarcinoma” was observed ($R = 0.808$) (see Appendix Table 4 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011). In addition, after running collinearity diagnostics, high variance inflation factor (>5) and low tolerance (<0.2) values were observed for both “Male” and “Adenocarcinoma.” Therefore, we omitted the “Male” variable from both multivariate regression models. Table 3 presents the full multivariate linear regression model. In data cut D, an additional 1 month of APFS should translate into 1.13 AOS months (95% CI 0.531–1.727), after adjustment for all other variables. Appendix Table 5 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011 shows the results from the logistic regression model, in which data cut D is the dependent variable. Studies with no crossover and reported balanced post-progression treatments seem to be of higher quality (higher Jadad scale) and less likely to have EGFR-TKI as investigated treatment.

Discussion

This study aimed to evaluate ORR and PFS as surrogate endpoints for OS in trials involving patients with advanced NSCLC receiving second- and further-line therapy. The impact of crossover and unbalanced post-progression treatments on surrogacy was assessed. Our findings show that crossover (per- and off-protocol) and unbalanced post-progression treatments may bias the association between the surrogate endpoints such as ORR or PFS and OS. When all trials were included in the analysis, the correlation coefficients of ORR and mPFS were 0.181 (95% CI 0.016–0.337) and 0.254 (95% CI 0.074–0.418), respectively, with OS. According to the IQWiG, this suggested that ORR and PFS are not valid surrogate endpoints for OS [4]. Different results are seen in analyses in which we included trials explicitly reporting balanced post-progression treatments and excluded trials with either per-protocol or off-protocol crossover, unbalanced post-progression treatments, or no information (data cut D). Both ORR and PFS had stronger associations with OS (ORR and OS: $R = 0.528$; 95% CI 0.081–0.798; PFS and OS: 0.778; 95% CI 0.475–0.916). Nevertheless, the upper limit of the correlation coefficient CI in both cases was higher than 0.7. Consequently, according to IQWiG recommendations, the validity of ORR and PFS as surrogate endpoints for OS is unclear [4]. In this case, the treatment effect of the surrogate in clinical trials needs to be statistically significant. This study has some limitations. We have not searched for unpublished studies. Nevertheless, we believe that we did not miss relevant studies for various reasons. First, information about crossover or post-progression treatments is mostly reported in full study reports. Thus, including, for example, an abstract database would not have added to our conclusion because all retrieved hits would have been excluded at the end. In addition, according to the funnel plots, publication bias is unlikely. Second, excluding studies from the main analysis on the basis of crossover or unbalanced post-progression treatment may have introduced selection bias. Nevertheless, the population characteristics of the trials included in different analyses appear to be comparable (Table 1), minimizing the risk of selection bias. Third, the populations, settings, and interventions included are heterogeneous. This might also affect the association between the surrogate and the true endpoints. On the basis of the multivariate model, however, the conclusion that crossover and unbalanced post-progression treatments may bias such relationship still holds. Only median age may have a minimal impact. Fourth, the definition and assessor of response is not consistent across studies and we did not account for that in our analysis. In previous endpoint literature, however, presented analyses were not stratified by response definition [6,10–13]. In addition, stratifying our analysis on the basis of response criteria would have impeded carrying out analysis in different data cuts.

Applying our methodology, that is, stratifying studies by crossover and post-progression treatments, for other surrogate endpoints in NSCLC (e.g., time to progression and duration of response), in different line settings (e.g., first-line NSCLC) or in other tumor types, is warranted. In other tumor types, it is likely that crossover and unbalanced post-progression treatments may
bias the relationship between surrogate endpoints and hard endpoints in these settings. Nevertheless, the extent of such likely bias needs to be further assessed.

Conclusions

Crossover and postprogression treatments may bias the quantitative relationship between surrogate endpoints (ORR/PFS) and OS. Therefore, the validity of ORR and PFS as surrogate endpoints for OS might be better estimated in trials that do not allow for crossover and that report balanced postprogression treatments. In second- and further-line therapy of advanced NSCLC, the validity of ORR and PFS as surrogate endpoints for OS is unclear, and a large effect size is needed to predict OS benefit with sufficient certainty. Trials that show a statistically significant treatment effect of 41% ORR or 4.2 mPFS months are expected to show a significant OS benefit with sufficient certainty. Further investigation of such methodology for other surrogate endpoints, in frontline therapy and in other tumor types and settings, is warranted.

Acknowledgment

We thank Andy Noble (Bioscript Science, Macclesfield, UK) for editorial support.

Source of financial support: This research was supported by Merck KgaA.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2017.07.011 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES