Evaluating surrogacy metrics and investigating approval decisions of progression-free survival (PFS) in metastatic renal cell cancer: a systematic review

K. R. Johnson1*, W. Liauw2 & M. N. D. Lassere2

Departments of 1Rheumatology; 2Oncology, St George Clinical School–University of New South Wales, Kogarah, Australia

Received 15 December 2013; revised 25 February 2014, 5 April 2014 and 30 May 2014; accepted 1 July 2014

Background: In metastatic renal cell cancer (mRCC) trials, progression-free survival (PFS) is increasingly used instead of overall survival (OS) as the approval end point. Unlike other solid tumors, there is no published demonstration of what PFS is needed across and by treatment class in mRCC. We determine this and evaluate drug approval decisions in mRCC targeted therapy.

Methods: We identified all randomized, controlled trials reporting PFS and OS in mRCC. Surrogacy metrics were the coefficient of determination and surrogate threshold effect (STE)—the PFS difference needed to predict, with 95% confidence, an OS difference. Data from regulatory commentaries, briefing documents and transcripts were extracted.

Results: No exclusively chemotherapy trial met criteria. Of 30 qualifying trials, 11 trials (13 comparisons) used targeted therapy. The all-trials and immunotherapy-only trials analysis failed to demonstrate a STE. The targeted trials, using the more conservative regression analysis demonstrated an STE of 3.9 months and an R² of 0.44. Crossover upon progression, control to active treatment, was common. Regulatory approval, accelerated or regular, labeling, interim analyses, and adjudication were context specific.

Conclusions: A new targeted therapy trial showing a PFS difference of 3.9 months can claim an OS benefit in mRCC. PFS surrogacy for OS in metastatic renal cell is not generalizable across all drug classes.

Key words: metastatic renal cell, surrogate marker, targeted therapy, surrogate threshold effect, progression-free survival

*Correspondence to: Dr Kent R. Johnson, Department of Rheumatology, St George Clinical School, University of New South Wales, St George Hospital, Belgrave St Kogarah, NSW 2217, Australia. Tel: +61-4-2862-1346; E-mail: kent.johnson@unsw.edu.au

© The Author 2014. Published by Oxford University Press on behalf of the European Society for Medical Oncology. All rights reserved. For permissions, please email: journals.permissions@oup.com.
Introduction

The generally poor response of metastatic renal cell cancer (mRCC) to conventional chemotherapy plus reports of spontaneous tumor regression led to immunotherapy [1], then targeted therapy trials with the first large targeted therapy trial, sunitinib [2], starting in 2003. Since 2005, four vascular endothelial growth factor (VEGF) receptor inhibitors, sorafenib (2005), sunitinib (2006) [3], pazopanib (2009) [4], and axitinib (2012) [5], two mammalian target of rapamycin (mTOR) inhibitors, temsirolimus (2007) [6], and everolimus (2009) [7], and one monoclonal antibody to VEGF, bevacizumab (2009) [8], have been marketed in the United States. During this time, there has been increasing interest in nonoverall survival (OS) trial end points for metastatic cancer, particularly progression-free survival (PFS), a composite of tumor progression and death. At the same time, there has been increasing use of ‘crossover’ designs, allowing control patients to switch to the active arm treatment (or to other nonprotocol ‘rescue’ treatments) at progression, and of blinded end point adjudication [1], allowing bias reduction in ascertaining progression. In a 2006 FDA workshop, the quantitative surrogacy relationship between PFS and OS in randomized, controlled trials of first-line advanced ovarian cancer [9] was presented. PFS/OS relationships have also been demonstrated in metastatic colon cancer [10, 11], fore-shadowing accelerated approval of panitumumab in second-line EGFR+ metastatic disease [12]. In 2010, FDA discussed the oncology review process and approvals from July 2005 to December 2007 in four common solid tumor settings, breast, colon, ovary, and lung [13, 14]. In mRCC PFS is increasingly interwoven in the approval process [15], although the PFS/OS surrogacy relationship has not been reported. In the present work, we demonstrate quantitatively how PFS performs as a surrogate of OS [16, 17] in trials of mRCC, and we qualitatively review PFS use targeted therapy approvals in this condition.

Methods

Search strategy

Part 1—quantitative review. A Medline search (OVID, 1946 to 15 January 2013) used the strategy: exp Carcinoma, Renal Cell/AND limit to (clinical trial, phase II or clinical trial, phase III or controlled clinical trial or randomized controlled trial) AND random$.mp. The ClinicalTrials.Gov and Cochrane Library databases were also searched. There was no language restriction. All citation titles and abstracts were examined by two reviewers (KJ, ML) to identify trials that met our inclusion criteria.

Part 2—qualitative review. Using Medline, FDA, and EMA regulatory websites, and ClinicalTrials.gov, we obtained reports, reviews, commentaries and regulatory meeting briefing documents and transcripts on PFS as an end point in mRCC. Secondary searching of all cited references from full-text publications was also conducted.

Inclusion/exclusion criteria and data extraction

Part 1—quantitative review. Inclusion criteria were full-text (not abstracts alone) publications reporting by-arm median PFS and OS, in validly randomized mRCC trials (but excluding randomized discontinuation designs), and evaluating nonsurgical drug therapy (chemotherapy, immunotherapy, or targeted therapy). Targeted therapy includes agents directed against antiangiogenic and molecular targets including mTOR and VEGF. Intention-to-treat data were extracted by two independent reviewers (KJ, ML) including PFS and OS results, treatment and control, trial size, trial conduct period, open/blind design, first or second-line therapy, crossover provisos, proportion with ECOG = 0, proportion with only one metastatic site, and Memorial Sloan-Kettering Cancer Center (MSKCC) scores.

Part 2—qualitative review. A single reviewer extracted from Medline, regulatory literature (KJ), and clinicaltrials.gov (ML) the following: (i) nature of primary and secondary end points, (ii) interim analyses prespecified and/or carried out with outcomes and alpha spending values, (iii) approval type (accelerated or regular), (iv) approval and/or carry over treatments) at progression, and of blinded end point adjudication [1], allowing bias reduction in ascertaining progression. In a 2006 FDA workshop, the quantitative surrogacy relationship between PFS and OS in randomized, controlled trials of first-line advanced ovarian cancer [9] was presented. PFS/OS relationships have also been demonstrated in metastatic colon cancer [10, 11], fore-shadowing accelerated approval of panitumumab in second-line EGFR+ metastatic disease [12]. In 2010, FDA discussed the oncology review process and approvals from July 2005 to December 2007 in four common solid tumor settings, breast, colon, ovary, and lung [13, 14]. In mRCC PFS is increasingly interwoven in the approval process [15], although the PFS/OS surrogacy relationship has not been reported. In the present work, we demonstrate quantitatively how PFS performs as a surrogate of OS [16, 17] in trials of mRCC, and we qualitatively review PFS use targeted therapy approvals in this condition.

Methods

Search strategy

Part 1—quantitative review. A Medline search (OVID, 1946 to 15 January 2013) used the strategy: exp Carcinoma, Renal Cell/AND limit to (clinical trial, phase II or clinical trial, phase III or controlled clinical trial or randomized controlled trial) AND random$.mp. The ClinicalTrials.Gov and Cochrane Library databases were also searched. There was no language restriction. All citation titles and abstracts were examined by two reviewers (KJ, ML) to identify trials that met our inclusion criteria.

Part 2—qualitative review. Using Medline, FDA, and EMA regulatory websites, and ClinicalTrials.gov, we obtained reports, reviews, commentaries and regulatory meeting briefing documents and transcripts on PFS as an end point in mRCC. Secondary searching of all cited references from full-text publications was also conducted.

Inclusion/exclusion criteria and data extraction

Part 1—quantitative review. Inclusion criteria were full-text (not abstracts alone) publications reporting by-arm median PFS...
OS relationship. We used Stata 12 for all analyses and the PRISMA statement [22] for reporting our results.

**Qualitative review.** This work was descriptive with no statistical methods other than simple tabulations.

**Results**

**Quantitative review**

Medline search yielded 318 citations of which 229 were discarded not meeting criteria after abstract review (Figure 1). Full texts of 89 remaining studies were examined, yielding 30 qualifying trials. The majority of excluded trials failed to report PFS and OS as many were conducted before PFS use was widespread. Small, short trials were rarely powered for OS and often did not report it. Two recent Cochrane reviews [23, 24] yielded no additional qualifying trials. No unpublished study was obtained. The 30 qualifying trials yielded 41 comparisons (Table 1), 11 targeted therapy trials (13 comparisons), and 19 immunotherapy trials (28 comparisons). All targeted therapy trials enrolling after 2000 used RECIST criteria (Version 1.0) [54]. While earlier trials tended to use the older WHO criteria. No trial of exclusively chemotherapy met criteria. One published report [38] was two trials with different treatments and different populations, one enrolling only patients with pulmonary metastases and one excluding such patients. mRCC trials generally excluded poorly functioning patients by ECOG status or equivalent measure.

The EIV targeted therapy regression model (Figure 2, upper graph) with reliability coefficient 0.9 showed satisfactory diagnostics (normality of error, no heteroskedasticity). A reliability coefficient of 0.9 indicates minimum PFS measurement error which, as noted, is conservative. The slope of the regression is 0.54 (P = 0.01) with an $R^2$ of 0.49 indicating that approximately half of the OS difference is explained by the PFS difference. The STE, the intersection of the lower 95% prediction line and the PFS axis, is 3.65 months. This means a future targeted therapy trial would need a PFS difference of at least 3.7 months to predict with 95% confidence an OS benefit. The model was not improved by adding terms for EGOG status, first/second-line...
Table 1. Design, clinical characteristics and results of 11 targeted therapy (T) and 19 immunotherapy (I) trials in metastatic renal cell carcinoma

| Author | I or T | Final pub year | Trial conduct period | Test arm treatment | Control arm treatment | First, second, line WHO = 0 Kar
Karnofskya | %ECOG | Blind | Size test arm | Size control arm | Median OS diff (months) | Median PFS diff (months) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Boccardo et al. [25]</td>
<td>I</td>
<td>1998</td>
<td>1993–1995</td>
<td>IL2 + IFN-α</td>
<td>IFN-α</td>
<td>1</td>
<td>0.77</td>
<td>0</td>
<td>22</td>
<td>22</td>
<td>−5.0</td>
<td>−1.5</td>
</tr>
<tr>
<td>Negrier et al. [26]</td>
<td>I</td>
<td>1998</td>
<td>1993–1995</td>
<td>IL2</td>
<td>IFN-α</td>
<td>1</td>
<td>0.89</td>
<td>0</td>
<td>22</td>
<td>22</td>
<td>10.0</td>
<td>4</td>
</tr>
<tr>
<td>Negrier et al. [26]</td>
<td>I</td>
<td>1998</td>
<td>1992–1995</td>
<td>IL2 + IFN-α</td>
<td>IL2</td>
<td>1</td>
<td>0.77</td>
<td>0</td>
<td>140</td>
<td>138</td>
<td>5.0</td>
<td>1</td>
</tr>
<tr>
<td>Negrier et al. [26]</td>
<td>I</td>
<td>1998</td>
<td>1992–1995</td>
<td>IL2 + IFN-α</td>
<td>IFN-α</td>
<td>1</td>
<td>0.77</td>
<td>0</td>
<td>140</td>
<td>147</td>
<td>4.0</td>
<td>1</td>
</tr>
<tr>
<td>MRCRC [27]</td>
<td>I</td>
<td>1999</td>
<td>1992–1997</td>
<td>IL2</td>
<td>MPA</td>
<td>1</td>
<td>0.26</td>
<td>0</td>
<td>167</td>
<td>168</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td>Pyrthonon et al. [28]</td>
<td>I</td>
<td>1999</td>
<td>1988–1994</td>
<td>IFN-α + vinblastine</td>
<td>Vinblastine</td>
<td>1</td>
<td>0.17</td>
<td>0</td>
<td>79</td>
<td>81</td>
<td>7.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Motzer et al. [29]</td>
<td>I</td>
<td>2000</td>
<td>1994–1996</td>
<td>Retinoic acid + IFN-α</td>
<td>IFN-α</td>
<td>1</td>
<td>0.61</td>
<td>0</td>
<td>61</td>
<td>59</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Negrier et al. [30]</td>
<td>I</td>
<td>2000</td>
<td>1995–1996</td>
<td>IL2</td>
<td>IL2 + IFN-α</td>
<td>1</td>
<td>0.425</td>
<td>0</td>
<td>61</td>
<td>70</td>
<td>0.0</td>
<td>1</td>
</tr>
<tr>
<td>Negrier et al. [30]</td>
<td>I</td>
<td>2000</td>
<td>1995–1996</td>
<td>IL2 + IFN-α</td>
<td>TAMOXIFEN</td>
<td>1(2)</td>
<td>NR</td>
<td>0</td>
<td>41</td>
<td>37</td>
<td>11.0</td>
<td>7</td>
</tr>
<tr>
<td>Dutcher et al. [32]</td>
<td>I</td>
<td>2003</td>
<td>1990–1992</td>
<td>IFN-α + IFN-γ</td>
<td>IFN-γ</td>
<td>1</td>
<td>0.34</td>
<td>0</td>
<td>49</td>
<td>39</td>
<td>3.9</td>
<td>1.5</td>
</tr>
<tr>
<td>Yang et al. [33]</td>
<td>I</td>
<td>2003</td>
<td>1998–2001</td>
<td>Low-dose bevacizumab</td>
<td>Placebo</td>
<td>2</td>
<td>0.78</td>
<td>1</td>
<td>37</td>
<td>40</td>
<td>3.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Atzpodien et al. [34]</td>
<td>I</td>
<td>2004</td>
<td>1995–1998</td>
<td>IL2 + IFN-α</td>
<td>IFN-α + vinblastine</td>
<td>1</td>
<td>NR</td>
<td>0</td>
<td>146</td>
<td>63</td>
<td>11.0</td>
<td>2</td>
</tr>
<tr>
<td>Aass et al. [35]</td>
<td>I</td>
<td>2004</td>
<td>1995–1998</td>
<td>IL2 + IFN-α + 5FU</td>
<td>Retinoic acid</td>
<td>1</td>
<td>0.57</td>
<td>0</td>
<td>159</td>
<td>161</td>
<td>4.1</td>
<td>1.9</td>
</tr>
<tr>
<td>Donskov, Denmark [36]</td>
<td>I</td>
<td>2005</td>
<td>2000–2001</td>
<td>IL2</td>
<td>IL2</td>
<td>1</td>
<td>0.91</td>
<td>0</td>
<td>33</td>
<td>30</td>
<td>6.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Donskov, England [36]</td>
<td>I</td>
<td>2005</td>
<td>1999–2001</td>
<td>IL2</td>
<td>IL2</td>
<td>1</td>
<td>0.77</td>
<td>0</td>
<td>21</td>
<td>20</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>McDermott et al. [37]</td>
<td>I</td>
<td>2005</td>
<td>1997–2000</td>
<td>High-dose IV-IL2</td>
<td>scIL2 + IFN-α</td>
<td>1</td>
<td>0.65</td>
<td>0</td>
<td>96</td>
<td>96</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Atzpodien et al. [38]</td>
<td>I</td>
<td>2006</td>
<td>NR</td>
<td>IL2 + IFN-α + retinoic acid + 5FU</td>
<td>IL2 + IFN-α</td>
<td>1</td>
<td>NR</td>
<td>0</td>
<td>65</td>
<td>78</td>
<td>−4.0</td>
<td>−1</td>
</tr>
<tr>
<td>Tannir et al. [39]</td>
<td>I</td>
<td>2006</td>
<td>2002–2003</td>
<td>Low-dose IFN</td>
<td>IFN</td>
<td>1</td>
<td>0.456</td>
<td>1</td>
<td>59</td>
<td>59</td>
<td>8.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Ebbinghaus et al. [40]</td>
<td>I</td>
<td>2007</td>
<td>2003–2004</td>
<td>Low-dose ABT-510</td>
<td>High-dose ABT-510</td>
<td>1</td>
<td>0.7</td>
<td>0</td>
<td>51</td>
<td>52</td>
<td>1.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Escudier et al. [41]</td>
<td>I</td>
<td>2007</td>
<td>2000–2002</td>
<td>Shark cartilage AE941</td>
<td>Placebo</td>
<td>2</td>
<td>0.527</td>
<td>0</td>
<td>151</td>
<td>151</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>Hudes et al. [42]</td>
<td>I</td>
<td>2007</td>
<td>2003–2005</td>
<td>Temsirolimus</td>
<td>IFN</td>
<td>1</td>
<td>NR</td>
<td>0</td>
<td>209</td>
<td>207</td>
<td>3.6</td>
<td>2.4</td>
</tr>
<tr>
<td>Negrier et al. [43]</td>
<td>I</td>
<td>2007</td>
<td>2003–2005</td>
<td>Temsirolimus + IFN</td>
<td>IFN</td>
<td>1</td>
<td>NR</td>
<td>1</td>
<td>210</td>
<td>207</td>
<td>1.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Negrier et al. [44]</td>
<td>I</td>
<td>2008</td>
<td>2000–2005</td>
<td>IVIL2 + IFN-α</td>
<td>IL2</td>
<td>1</td>
<td>0.35</td>
<td>1</td>
<td>122</td>
<td>125</td>
<td>1.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Escudier et al.—TARGET [45]</td>
<td>I</td>
<td>2009</td>
<td>2003–2005</td>
<td>Sorafenib</td>
<td>Placebo</td>
<td>2</td>
<td>0.48</td>
<td>1</td>
<td>451</td>
<td>452</td>
<td>2.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Motzer et al. [46]</td>
<td>I</td>
<td>2009</td>
<td>2004–2005</td>
<td>Sunitinib</td>
<td>IFN-α</td>
<td>1</td>
<td>0.62</td>
<td>0</td>
<td>375</td>
<td>375</td>
<td>4.6</td>
<td>6</td>
</tr>
<tr>
<td>Study</td>
<td>Duration</td>
<td>Comparator</td>
<td>Intervention</td>
<td>Number</td>
<td>Objective</td>
<td>Result</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>------------</td>
<td>--------------</td>
<td>--------</td>
<td>-----------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escudier et al. [47]</td>
<td>2004-2006</td>
<td>Bevacizumab + IFN-α</td>
<td>Carnitine + plitidepsin</td>
<td>327</td>
<td>0</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escudier et al. [47]</td>
<td>2004-2006</td>
<td>Bevacizumab + IFN-α</td>
<td>Carnitine + plitidepsin</td>
<td>327</td>
<td>0</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gore et al. [48]</td>
<td>2001-2006</td>
<td>IL2 + FU + IFN-α</td>
<td>Carnitine + plitidepsin</td>
<td>90206</td>
<td>0</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motzer et al. [50]</td>
<td>2001-2006</td>
<td>Everolimus</td>
<td>Placebo</td>
<td>151</td>
<td>0</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu et al. [52]</td>
<td>2005-2008</td>
<td>Autologous CIK</td>
<td>IL2 + IFN-α</td>
<td>74</td>
<td>0</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- Usual restriction was ECOG/WHO ≤ 1; the equivalent in the Karnofsky scale is 90% or more.
- Negrier [41] used a factorial design with six comparisons, adjusted per Bonferroni for multiple comparisons (see text, Statistical Analysis).
- This notation (2) indicates that only a small minority were second-line patients.
- This trial of 302 patients did not note number of patients per arm, so 151 was used for both arms.
- I, included in immunotherapy analysis; T, included in targeted therapy analysis; IFN-α, interferon-α; IFN-γ, interferon-γ; IL2, interleukin 2; CIK, cytokine-induced killer; FU, fluorouracil; MPA, medroxyprogesterone acetate; NR, not reported.

**Qualitative Review**

The evaluation of regulatory approvals of targeted agents in mRCC is summarized in Tables 2 and 3. For reasons of consistency, we have focused on the US experience in this report because of the longer history of available FDA information including transcripts of advisory meetings. However, recent publications from EMA indicate similar considerations in its approvals. The tables are structured chronologically and include trial design and efficacy results available in the public domain. They are not exhaustive because some aspects of risk-benefit, e.g., the full safety database, are not included. However, the table includes FDA approval type (accelerated or regular), indication, approval year, trial population, active and comparator groups, primary end point, interim analyses, alpha spending, blinding, crossover/rescue therapy, adjudication provisions, and PFS and OS trial results. As can be seen, of seven targeted therapy approvals only sunitinib was given accelerated approval based on durable partial response rates, 25% and 36%, in PFS and OS trial results. The tables are structured chronologically and include trials of mRCC that were more heterogeneous in design and outcome than targeted therapy trials. The failure of regression diagnostics means no STE conclusion can be made from the model. The combined targeted and immunotherapy trials are shown in Figure 2, lower graph, but visual inspection adjusting for the different vertical axis scales reveals that the immunotherapy model slope is quite different from the targeted therapy model slope, suggesting that combining them is inappropriate.

**Discussion**

We have evaluated PFS as a surrogate for OS in trials of mRCC using targeted and immunotherapy and have found a STE for therapy, percent MSKCC favorable disease, or involvement of only one metastatic site. The targeted therapy ordinary linear regression model was similar: slope of 0.48 (P = 0.013), adjusted R² of 0.44, and STE of 3.9 months, indicating that the ordinary linear regression is here more conservative than the EIV regression.

In contrast, the immunotherapy trial model (Figure 2, middle graph) showed unsatisfactory diagnostics with residual nonnormality and heteroskedasticity, not corrected by fractional polynomial regression transformation with up to three additional terms or inclusion of omitted prognostic covariates (first/second line, performance status, proportion with only one metastatic site). Clinically, immunotherapy trials were more heterogeneous in design and outcome than targeted therapy trials. The failure of regression diagnostics means no STE conclusion can be made from the model. The combined targeted and immunotherapy trials are shown in Figure 2, lower graph, but visual inspection adjusting for the different vertical axis scales reveals that the immunotherapy model slope is quite different from the targeted therapy model slope, suggesting that combining them is inappropriate.
By-trial weighted least squares errors-in-variables regression of median overall survival (OS) difference and median progression-free survival (PFS) difference for targeted therapy trials weighted by trial size (squares, upper graph), immunotherapy trials (circles, middle graph) and all trials (lower graph) assuming minimal measurement error (reliability coefficient = 0.9). The upper and lower solid bold lines are the upper and lower 95% prediction limits; the dashed inner lines are the 95% confidence limits; the dot-dash centre line is the mean regression line. The Surrogate Threshold Effect (STE) is the intersection of the lower 95% prediction line and the PFS axis and indicates the median PFS difference needed to impute a median overall survival benefit in a new trial. Only the targeted therapy (upper graph) STE is valid; the other two models fail in their diagnostics (see text).

Our model demonstrates that a new trial needs a PFS difference of at least 3.65 months in order to predict with 95% confidence a benefit in OS difference. As can be seen from Figure 2 (upper), a trial with a PFS difference of 3.65 months would indicate an approximate 2.5-month OS difference. This 2.5-month OS gain is for a mixed population: (i) crossover (or other active rescue treatment) patients, and (ii) noncrossover patients, i.e. still on their randomly assigned treatment. The demonstration would be falsified (i.e. rendered a false-positive inference) only if the crossover population, in reality, had a better OS benefit than the population still on randomized treatment, an assertion that, from all evidence, is exceedingly unlikely.

It is important to note that correlation alone cannot justify a surrogate marker argument. This was first noted by Fleming and DeMets [57] and also by Burzykowski et al. [16] and Lassere [19]. Surrogacy also requires that the effect of an intervention on the surrogate predict the effect on the true outcome, a stronger condition than correlation. A very recent study [58] of individual patient data from two of the targeted therapy trials [48, 51] reported correlations and associations using Cox proportional hazards models of OS outcomes versus 3- and 6-month landmark (binary) PFS outcomes with treatment assignment as a covariate. They demonstrated statistically significant Kendall tau coefficient measures of association of 0.53 and 0.50 and statistically significant adjusted association measures, and they concluded that this suggested that PFS may be used as a surrogate end point for OS. This study provides patient-level measures, but in the opinion noted in the accompanying editorial [59], it does not provide evidence that the surrogate fully captured the effect of the intervention on the end point. This is Prentice’s fourth criterion; however, it is a criterion that is very difficult to assess because it is formulated in terms of an equivalence setting [60]. Others, therefore, have offered alternatives to the Prentice criteria. These alternatives are founded on prediction in a multitrial framework [16, 18]. In this approach, evidence for statistical surrogacy is strengthened by two trial-level statistics. The first statistic is a relative value, the meta-analytic coefficient of determination of the regression between trial-level effects of treatment on both the surrogate and clinical end point. This is the $R^2$ trial. The second statistic is an absolute value, here the STE, which incorporates both the slope of the PFS–OS regression relationship, its intercept, its dispersion, and the level of uncertainty in a prediction model. The role of the STE is to
### Table 2. Trial population and design of FDA approved metastatic renal cell carcinoma targeted drugs

<table>
<thead>
<tr>
<th>Name, approval type, sources: 1. General published literature; 2. FDA authored published literature; 3. Dedicated FDA advisory committee meeting material and transcript; 4. Advisory committee material from meetings for other drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indication</strong></td>
</tr>
<tr>
<td>Sorafenib</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Sunitinib</td>
</tr>
<tr>
<td>Temsirolimus</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Bevacizumab + interferon</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Pazopanib</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Axitinib</td>
</tr>
</tbody>
</table>

*Sunitinib was given accelerated approval based on durable partial response rates (25% and 36%) in two single-arm studies because these rates were much higher than those seen in TARGET (2%, sorafenib, 0%, placebo) [24]. The 750 patient trial was the basis for subsequent conversion to regular approval.

†It is unclear if there was a blinded adjudication committee throughout AVOREN. The primary source [48] makes no mention of such, and Summers [8] notes that in 2006, 2 years into AVOREN, that a statistically robust and clinically important effect on PFS that was confirmed by an independent review committee masked to treatment assignment could serve as a basis for label expansion. Similarly, Summers [8] notes that the CALGB trial ‘was similar in design’ to AVOREN.

IFN, interferon-α; KPS, Karnofsky performance status; an alternative to ECOG/WHO where KPS100 and KPS90 are approximately equal to ECOG, 0; MSKCC, Memorial Sloan-Kettering Cancer Center; NR, not reported.
<table>
<thead>
<tr>
<th>Drug</th>
<th>Primary endpoint (also see Results)</th>
<th>No. of interim analyses</th>
<th>Analysis number and results, median (months): test versus control, HR (95% CI), P-value, α needed (where applicable), trial amendment, trial termination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorafenib TARGET</td>
<td>OS</td>
<td>1 2</td>
<td>5.5 versus 2.8 months 0.44 (0.35–0.55), P &lt; 0.001 α = 0.01 Trial amended to allow crossover (1) NR versus 14.7 months 0.72 (0.54–0.94), P = 0.015 α = 0.0094 Precrossover (2) 19.3 versus 15.9 months 0.77 (0.63–0.95), P = 0.015 α = ‘smaller’</td>
</tr>
<tr>
<td>Sunitinib</td>
<td>PFS</td>
<td>3 3</td>
<td>(2) 11 versus 5 months 0.42 (0.32–0.54), P &lt; 0.001 α: NR Trial amended to allow crossover (2) NR versus NR HR = 0.65 (0.45–0.94), P = 0.02 α = not met</td>
</tr>
<tr>
<td>Temsirolimus</td>
<td>OS</td>
<td>0 2</td>
<td>5.5 versus 3.1 months HR: 0.66, P = 0.001 (2) 10.9 versus 7.3 months 0.73 (0.58–0.92), P = 0.008 α = 0.01325 Trial terminated as alpha met (2) 14.8 versus 14.4 months 0.87 (0.65–1.15), P = 0.162</td>
</tr>
<tr>
<td>Everolimus</td>
<td>PFS</td>
<td>2 0</td>
<td>(2) 4.9 versus 1.9 months 0.33 (0.25–0.43), P &lt; 0.001 α = 0.0057 Trial terminated as alpha met (2) 14.8 versus 14.4 months 0.87 (0.65–1.15), P = 0.162</td>
</tr>
<tr>
<td>Bevacizumab + IFN AVOREN</td>
<td>OS</td>
<td>1 1</td>
<td>10.2 versus 5.4 months 0.63 (0.52–0.75), P = 0.0001 α: not reported NR versus 19.8 months 0.79 (0.62–1.02), P = 0.067 α = 0.0056 Trial terminated as alpha met</td>
</tr>
<tr>
<td>Bevacizumab + IFN CALGB 90206</td>
<td>OS</td>
<td>7 7</td>
<td>(#6) 8.5 versus 5.2 months 0.72 (0.61–0.83), P &lt; 0.0001 α: not reported (6) 18.3 versus 17.4 months 0.86 (0.73–1.01), P = 0.69 α: NR Trial terminated accrual met</td>
</tr>
<tr>
<td>Pazopanib</td>
<td>PFS</td>
<td>0 1</td>
<td>0.73 (0.53–1.00), P = 0.02 α = 0.004 [per FDA] 9.2 versus 4.2 months 0.46 (0.34–0.62), P &lt; 0.0001 To be reported when data mature</td>
</tr>
<tr>
<td>Axitinib</td>
<td>PFS</td>
<td>0 0</td>
<td>6.7 versus 4.7 months 0.67 (0.54–0.81), P &lt; 0.0001 α = 0.023 20.1 versus 19.2 months 0.97 (0.80–1.17), NS [per US label] To be reported when data mature</td>
</tr>
</tbody>
</table>

HR, hazard ratio; NR, not reported.
express the multtrial regression of OS differences versus PFS differences. The failure of using simple correlational analysis such as the Pearson correlation is demonstrated by an evaluation of our results. The calculated Pearson correlations were 0.67 for the targeted therapy trials and 0.54 for the immunoimmunotherapy trials, yet an STE could only be demonstrated for the targeted therapy trials. As we did not have patient-level data, we were unable to determine the multtrial equivalent of the adjusted association, the $R^2_{adj}$ individual. The $R^2_{adj}$ individual is the coefficient of determination of the patient-level association between both end points. However, Halabi et al. [58] found statistically significant adjusted associations of 3.145 and 2.80 in the targeted therapy trials CALGB 90206 [51] and AVOREN [48], respectively, providing further support for our trial-level findings. Nonetheless, there are also substantive arguments for surrogacy in addition to statistical arguments. Attempts to capture all relevant aspects of surrogacy can be found in the frameworks described by Boissel et al. [61], Temple et al. [62], Bucher et al. [63], De Gruttola et al. [64], and Lassere and co-workers [19, 20, 65].

We compared our review with two comprehensive Cochrane reviews [23, 24], and confirmed, where available, end points via clinicaltrials.gov. Of 23 targeted therapy trials in the Cochrane review, 14 were reported as providing PFS and OS; however, on review, 2 of these reported time to progression (TTP) rather than PFS, 1 reported only proportion progressed, and 2 did not report OS by arm. The remaining Cochrane targeted therapy trials were captured in our analysis, along with three additional targeted trials [41, 47, 53]. There were insufficient trials reporting TTP to model TTP and OS as has been done in other solid tumors [66]. We found no attempt to model quality of life in mRCC or any other solid tumor, probably because it is very difficult to reliably measure in cancer trials.

There were no exclusively chemotherapy trials that met our criteria. The immunotherapy model failed to meet necessary regression assumptions. Additionally, immunotherapy, being, at best, modestly efficacious and at times seriously toxic, will under-power any analysis. In the Cochrane immunotherapy review [24] evaluating remission rate difference versus OS difference, the authors did not present a formal regression model with validating diagnostics, but they noted that a number of studies showed discordant findings, improved remission but no survival benefit. Of 57 trials reported in the Cochrane review, 37 reported both remission and survival while 20 reported only remission. Only five reported quality of life.

To our knowledge, this is the first report of surrogacy metrics in mRCC that includes the STE. In an abstract [67] of PFS and TTP versus OS in mRCC trials, a significant slope and adjusted $R^2$ of 0.46 were found but no regression diagnostics or prediction bands were reported. This study combined targeted and immunotherapy trials which our data suggest is not valid, and it combined PFS with TTP and randomized with nonrandomized trials, both problematic. In an observational study, Heng et al. [68] reported on mRCC cohorts with consecutive enrollment treated with targeted therapy, finding a moderate correlation of PFS and OS ($R^2 = 0.44$). The STE model has been applied in other solid tumor settings such as advajlut colon cancer, metastatic colon cancer, metastatic nonsmall-cell lung cancer, breast cancer, and metastatic ovarian cancer [16, 66, 69, 70], and in vascular medicine where there is general agreement on surrogacy: blood pressure with all drug classes [19] and LDL cholesterol with statins [71]. As PFS is a composite end point capturing both progressions and deaths, its interpretation and clinical use will vary depending on the proportion of the events qualifying for the end point. If most events were death, PFS would offer little over assessing OS directly. However, generally, most mRCC patients progress before they die, and this finding is likely to result in most PFS events being progressions rather than deaths although trial reports rarely specifically reported these data.

The difference in the pharmacologic class models, targeted therapy, and immunotherapy goes to the heart of the STE concept. Beyond biologic plausibility, pharmacologic class, comparator treatment, background and rescue therapy, outcome assessment frequency, clinical heterogeneity (previous therapy, prognostic factors), treatment toxicity, differential drop-out (missing data or crossover), and measurement error all may influence the surrogate–outcome relationship. The driver of the surrogate–outcome relationship is context. Some STEs apply across all available pharmacologic drug classes, such as blood pressure [19], whereas others apply to only one drug class such as LDL cholesterol and statins [71]. Additionally, the STE does not substitute for individual patient-level data, although early data (unpublished observations) using hierarchical linear mixed model simulation suggests that patient-level data slightly decreases the magnitude of the STE. Therefore, our trial-level STEs may be conservative; that is, the PFS needed to predict an OS benefit could be somewhat <3.7 months. Furthermore, other important clinical dimensions impact the persuasiveness of evidence based on surrogate metrics [19, 62, 65].

From the perspective of drug development, issues with PFS usage include blinded independent central review, measurement error, image versus symptom-defined progression, asymmetric visit schedules, and bias due to differential reporting and to informative censoring in the presence of insufficient blinding [72–75]. In the eight large trials used for registrational purposes (Table 2), five were double-blind and all used assessed patients every 6–12 weeks for both arms for at least the first 6 months of the trial. Statistical issues include trial amendments allowing crossover/rescue therapy or termination based on interim analysis of a secondary end point. Dual track strategies (accelerated or regular approval) complicate approval strategy [55]. From FDA transcripts and briefing documents, we found that PFS could sometimes be considered a clinical benefit it its own right, without needing a surrogacy argument—improvements in the primary end point of PFS must be both clinically and statistically significant [76], and, from our past approvals from a regulatory perspective, PFS of a sufficient magnitude would be of clinical benefit [77]. However, ‘clinically significant’ and ‘sufficient magnitude’ were not quantified.

Protocol details such as interim analyses and crossover/rescue therapy are important. In sorafenib (TARGET), the interim PFS effect met its needed alpha but the primary OS end point had not, yet the trial was amended to allow crossover. In the sunitinib and bevacizumab (AVOREN) trials, the needed PFS alpha was not reported and no interim OS met the prespecified alpha, yet the trials were amended to allow crossover. No trial amended to allow crossover based on interim PFS results demonstrated a significant OS benefit at completion. Consequently, except for temsirolimus, uncertainty remains regarding whether a survival benefit exists. Failure to show an OS benefit could result from crossover/rescue
therapy attenuating a true benefit, but this failure would also result if the active drug were less effective or even deleterious in progressed patients. Some trials disallowed crossover [78], but availability of off-protocol therapy results in the same bias. There have been attempts to adjust for this bias in the analysis [50], but none of these approaches has yet been validated [79, 80]. Some clinicians argue that crossover is needed because the test drug cannot be denied to patients once they progress; however, other clinicians are concerned that crossover designs conflict with the need for unbiased OS assessment. ‘[O]ne should not ask patients to participate in a trial that cannot meet its scientific objectives’ [81]. One solution to this quandary is a design that incorporates a second randomization for control arm patients once they have progressed [82]. There is not, as yet, any direct demonstration that crossover confers an OS benefit.

Very recently, another mRCC targeted agent, tivozanib, failed to get FDA approval (http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM350075.pdf; http://www.fda.gov/downloads/advisorycommittees/committees meetingmaterials/drugs/oncologicdrugsadvisorycommittee/ucm359161.pdf). Similar to past advice [76], the FDA briefing document for the advisory meeting noted that during design discussions with the sponsor the FDA stated that ‘a substantial, robust improvement in PFS that is clinically meaningful and statistically persuasive may be considered for regulatory decision’ and that ‘a statistically significant improvement in OS is not required for regulatory approval, but a prespecified OS analysis plan is still helpful in the regulatory decision making process’ (http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM350075.pdf). The tivozanib submission consisted of a single open-label phase III study comparing tivozanib with sorafenib in mRCC patients with no prior targeted therapy. It was conducted primarily in Central and Eastern Europe. Crossover was permitted and occurred overwhelmingly from sorafenib to tivozanib. The trial found statistically significant improvement in PFS, 9.1 versus 11.9 months (HR = 0.80, 95% CI = 0.70–0.93) and non-significant worsening in OS, 28.2 versus 39.3 months (HR = 1.25, 95% CI = 0.95–1.64). This is a PFS difference of 2.8 which does not meet the STE criterion.

disclosure
The authors have declared no conflicts of interest.

references


