
Article:
Ciani, O., Cantrell, A. orcid.org/0000-0003-0040-9853, Davis, S. orcid.org/0000-0002-6609-4287 et al. (6 more authors) (2014) Validation of Surrogate Endpoints in Advanced Solid Tumors: Systematic Review of Statistical Methods, Results, and Implications for Policy Makers. International Journal of Technology Assessment in Health Care, 30 (3). pp. 312-324. ISSN 0266-4623

https://doi.org/10.1017/S0266462314000300

Supplementary technical note -

Statistical methods to assess the association between surrogate and final endpoints

Methods to examine observation-level association

Seven papers reported the relationship between median PFS or TTP and median OS using aggregate data.[1-7] Three treated both surrogate endpoints as different,[2,4,6] with one performing separate analyses for PFS, TTP and a composite measure including both PFS and TTP.[4] Two studies also examined the relationship between OS and post-progression survival (PPS), which was defined as the difference between median OS and median PFS or TTP.[3,7] In order to assess the correlation between the surrogate and final endpoints, five papers reported Spearman's ρ correlation coefficients,[1,2,4,5,7] one reported the Pearson product-moment correlation coefficient,[4] and three the coefficient of determination (R^2) or regression parameters derived from a linear regression analysis.[1,3,6] Statistical analyses were often weighted by trial size. Three papers included a variety of first-line treatments,[1-3] and one included only second- or third-line treatments.[7] In studies that included patients at different treatment lines, such line was a stratification factor in multivariate analyses (see Supplementary Table 3).

Seven IPD meta-analyses estimated 'individual-level' surrogacy between PFS or TTP and OS,[8-14] with the last two distinguishing between TTP and PFS or only considering TTP (on the log scale). Correlation between the surrogate and final endpoints at the individual-level was expressed through Spearman's ρ[8,9] or Pearson's[13] correlation coefficients, whereas two studies considered the patient-level agreement between PFS and OS at different time points,[10,11] with the latter study reporting a Kappa statistic to summarise the amount of agreement beyond that expected by chance alone. Individual-level correlation coefficients were derived from random-effects linear models of the association between normally distributed endpoints.[10,15] For failure-time endpoints,[8,12] Kendall's τ was used as a measure of the association between the surrogate and final endpoints, modelled through Hougaard's or Clayton' bivariate copula models. Landmark analysis[16] was used in six papers to assess the prognostic impact of being alive and progression-free at various timepoints on future survival.[11,17-21] In the landmark analysis, multivariate Cox proportional hazards models were constructed for OS and these were stratified by progression-free status at consecutive times. HRs

were reported for survival in patients who were alive and progression-free at these timepoints compared with those who were not. Three of the models were stratified by trial protocol,[17,18,21] while one reported separate analyses for each trial and a combined analysis adjusted for study protocol.[20] Two papers assessed the Kendall's τ rank correlation coefficient for bivariate censored data,[18,19] while Heng and colleagues[19] also assessed the correlation between PFS and OS using the Fleischer model.[22] Mandrekar and colleagues[21] and Foster and colleagues[17] evaluated model discrimination using the concordance index (c-index), which computes the probability that, for a pair of randomly chosen comparable patients, the patient with the lower risk prediction (e.g., progression-free at 3 months) will experience an event (e.g., death) before the higher risk patient (e.g., progressed before 3 months). A completely random prediction would have a c-index of 0·5, and perfect correlation will produce a c-index of 1·0.[21] Buyse et al.[8] and Halabi et al.[18] performed a validation procedure of their estimated models by dividing their samples into a training and a testing set.

Methods to examine treatment-level association

Fourteen studies examined the relationship between the treatment effect on PFS or TTP and the treatment effect on OS based on aggregate data.[2,4,5,23-33] Treatment effect was defined in several ways: absolute difference in medians of time-to-event endpoints,[2,24,30,33] proportional increase in medians of such endpoints,[25,26,31] or HRs.[4,5,23,27,28,30,32] One paper defined the treatment effect as the HR minus unity,[29] and another examined the percent risk reduction based on the HR.[2] Some authors transformed the HR onto a log scale for the linear regression,[23,28,33] and most of them defined the HR as the ratio of the median time-to-event between trial arms,[4,23,27-29] which implicitly assumes that the underlying distribution of event-free survival is exponential, although no justification was given for this assumption. The studies handled trials with more than two arms in a variety of ways. Most included multiple comparisons from the same trial as multiple points in the analysis without accounting for the correlations between them or the double-counting in terms of the sample size.[5,23,25,27,30,33] Linear regression analyses were the most common methods used to assess the relationship between treatment effect on PFS or TTP and treatment effect on OS based on summary data from multiple RCTs.[2,4,23,24,27-30,32,33] All but two reported that the regression analyses were weighted according to trial size.[2,30] Two studies did not force the intercept of the regression to zero,[2,29] although both considered and discounted a non-zero intercept in exploratory analyses. One study explored the

possibility of a nonlinear regression by adding quadratic terms.[29] One study[33] assessed the possibility of publication bias using funnel plot and Egger's test.[34] One study examined residual versus predicted plots and undertook diagnostic tests for normality and heteroscedasticity (non-constant error variance) to assess consistency with the assumptions of linear regression,[24] and another study evaluated the normality assumption and presence of outliers or influential points using diagnostic tests and plots.[4] Several authors used multivariate analysis to explore whether any other factors were significant predictors of treatment effect on OS (see Supplementary Table 3).[4,24,27,33] A 'leave-one-out' cross-validation to predict the OS HR from the PFS HR for each trial using a regression fitted to all the remaining trials was performed in two studies.[4,28] Other metrics used to ev[1]aluate trial-level surrogacy were the surrogate threshold effect (STE),[24] the Spearman's ρ,[2,4,5,25,26,31] or Pearson's correlation coefficients,[4,29,32,33] Kappa test for agreement,[28,29] or hypothesis sign test.[25,26,31] One paper built a receiver operating characteristics curve, a graphical display of the trade-off between sensitivity and specificity at various magnitudes of treatment effect for PFS, to assess whether the candidate surrogate endpoint is predictive of a clinically meaningful treatment effect in OS.[4]

Seven IPD meta-analyses reported estimates of the association between treatment effects on the surrogate and final endpoints.[8-10,12,14,17,35] Within the meta-analytic framework, trial-level surrogacy must be based on results from several randomized trials.[36] However, when an insufficient number of trials are available to conduct a meta-analysis, it is possible to break the results of large trials down into smaller units of analysis,[34] such as study centres. This expedient was used in four of the included studies.[10,12,14,17] Most of these studies expressed treatment effect as HRs for PFS, TTP and OS on the log scale,[8-10,17] while one study considered the absolute difference on TTP and OS on the log scale.[14] For the evaluation of the surrogate endpoints on the basis of IPD, the authors used joint models of the surrogate and the final endpoint as continuous bivariate normally distributed,[14] or time-to-event variables. Burzykowski and colleagues[9] used copula models, either Clayton's or Hougaard's types, to estimate trial- or centre-specific treatment effects on PFS or TTP and OS. Variations proposed to overcome statistical challenges for the computation and definition of the correlation coefficients in different situations have been discussed elsewhere.[36,37] In one paper, the regression was validated by using it to predict OS treatment effects from PFS treatment effects in three validation trials.[8] Burzykowski and Buyse[35] introduced the concept of STE for the first time and reported the minimum

HR required for PFS in order to observe a significant treatment benefit on OS in the context of advanced ovarian and colorectal cancers.

Supplementary Table 1 Biomarker-Surrogacy Evaluation Schema (BSES3)[§]

Biomarker-surrogate domains	
Study design	0 Biological plausibility and lower quality clinical studies 1 Rank 0 and at least 2 good quality prospective observational cohort studies measuring the surrogate and the target outcomes 2 Rank 1 and at least 2 high quality adequately powered RCTs measuring the surrogate and the target outcomes 3 Rank 1 and at least 5 high quality adequately powered RCTs measuring the surrogate and the target outcomes
Target outcome	0 Target is reversible disease-centred biomarker of harm 1 Target is irreversible disease-centred biomarker of harm 2 Target is patient-centred endpoint of reversible organ morbidity or clinical burden of disease or clinical harm 3 Target is patient-centred endpoint of irreversible organ morbidity or clinical burden of disease or severe irreversible clinical harm or death
Statistical evaluation of the biomarker-surrogate vs. target outcome	0 Poor: Does not meet the criteria for Rank 1 1 Fair: RCT $R^2_{trial} \geq 0.2$ AND STEP* ≥ 0.1 OR cohort data $R^2_{ind} \geq 0.4$ 2 Good: RCT $R^2_{trial} \geq 0.4$ AND STEP ≥ 0.2 AND $R^2_{ind} \geq 0.4$ 3 Excellent: RCT $R^2_{trial} \geq 0.6$ AND STEP ≥ 0.3 AND $R^2_{ind} \geq 0.6$**
Generalisability: clinical evidence across different risk populations and pharmacologic evidence across different drug-class mechanisms	0 No clinical or pharmacologic evidence 1 Clinical OR pharmacologic evidence 2 Clinical AND pharmacologic evidence 3 Consistent Clinical RCT AND pharmacologic RCT evidence
Level of evidence of surrogate endpoint multidimensional validity	
12	Level A
11 – 9	Level B+, B, B-
8 – 6	Level C+, C, C-, D+, D, D-
5 – 3	Level D+, D, D-, E+, E, E-
2 – 0	Level E+, E, E-, F+, F, F-

[§]Adapted from Lassere MN, Johnson KR, Schiff M, Rees D. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? An analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (STE) and the Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (BSES). *BMC Medical Research Methodology* 2012 Mar 12; **12**:27. doi: 10.1186/1471-2288-12-27.

* STEP is defined as the proportion of the total range of the surrogate that is equal or larger than the STE

**Without data subdivision. Some analyses with few trials subdivide into centres to increase the number of data points

Supplementary Table 2 Detailed characteristics of included meta-analyses, summary of statistical methods used and results

First author and year	Tumour type	Study identification	Inclusion criteria	N. of studies (patients)	Surrogate and final outcome relationships analysed	Statistical methods used to assess surrogate and final outcome association	Results
Summary data from trials							
Louvet et al. 2001[1]	mCRC	Not stated	Phase III studies of first line treatment reported between 1990 and 2000, >100 patients per study arm	29 13,498	Median PFS and median OS for individual trial arms	Spearman ρ correlation coefficient Linear regression	ρ =0.481, p <0.0001 OS (months) = 0.68 x PFS (months) + 8.74
Hackshaw et al. 2005[23]	mBC	Systematic search (Medline 1966-2005)	RCT comparing FAC or FEC with one or more first-line combination therapies	42 (9,163)	HR for TTP and OS (HR defined as ratio of median survival)	Linear regression on log-log scale weighted by sample size	$Log_{10}\ HR_{TTP} = 0.0135+0.5082 \times log_{10}\ HR_{OS}$ (p <0.001, R^2 = 56%, s.e. = 0.0928)
Johnson et al. 2006[24]	mCRC, mNSCLC	Systematic search	RCTs of first-line treatment	CRC: 146 (35,557) NSCLC: 191 (44,125)	Difference in median TTP and median OS	Linear regression weighted by trial size (multivariate analysis used to explore other potential predictive factors) STE for various trial sizes	mCRC: R^2 = 0.33; p <0.0001 OS = –0.002 + 0.0961 x TTP mNSCLC: R^2 = 0.19; p =0.0003 OS = 0.189 + 0.616 x TTP mCRC: 3.3 months mNSCLC: 3.2 months for trials of 250 patients
Tang et al. 2007[2]	mCRC	Systematic search	Randomised trials of first-line treatment published between 1990 and 2005, >100 patients per arm, mature data on OS and either TTP or PFS	39 (18,668)	Median PFS/TTP and OS Differences (Δ) in median OS, PFS and TTP HRs PFS and OS	Nonparametric Spearman rank correlation Linear regression (through origin) analysis	Median PFS and OS: ρ = 0.79 (95% CI, 0.65 to 0.87), p <0.000001 Median TTP and OS: ρ = 0.24 (95% CI,-0.13 to 0.55), p =0.21 ΔPFS and ΔOS : ρ = 0.74 (95% CI, 0.47 to 0.88), p = 0.00004 Slope = 1.02 (s.e. = 0.16), R^2=0.65 ΔTTP and ΔOS : ρ = 0.52 (95% CI, 0.004 to 0.81), p = 0.05 HR_{PFS} and HR_{OS}: Slope = 0.54 (s.e. = 0.10)
Bowater et al. 2008[25]	mBC, mCRC, HRP mNSCLC	Systematic search for reviews of RCT	RCTs published in English between 1990 and 2007 comparing two different chemotherapy treatments	BC: 33 (NS) CRC: 38 (NS) HRP: 23 (NS) NSCLC: 13	Gain (%) in median TTP and in post-progression survival (PPS) (PPS= median OS – median TTP)	Spearman's correlation Hypothesis (sign) test for proportion of trials with a) PPS%gain < TTP%gain, b) PPS%gain <0.5TTP%gain	ρ was non-significant at 10% level in all four disease areas a) p <0.001 for all four disease areas

				(NS)			b) p <0.005 for colorectal and p<0.001 for three other disease areas
Hotta et al. 2009[27]	Advanced or mNSCLC	Systematic search	Phase III trials of first-line therapies published between 1994 and 2006	54 (23,457)	Ratio of medians TTP and MST	Linear regression on ratios of medians TTP and MST Multivariate linear regression on on ratios of medians TTP and MST (weighted by trial size) incorporating 6 other factors	R^2 = 0.33, p <0.01 Multivariate analysis (R^2 = 0.41) gave regression coefficient of 0.32 (p <0.01) for TTP and no other factor was significant
Miksad et al. 2008[28]	Advanced breast (some locally advanced included)	Systematic search	RCTs published in English of anthracyclines and taxanes	31 (4,323)	HR for PFS and OS estimated by calculating the median OS and PFS ratios for each pair of trials arms	Kappa tests for agreement in direction of effects (HR) Fixed effects linear regression for LogHR (weighted by sample size)	Anthracyclines: Kappa = 0.71 (95% CI, 0.36 to 1.00, p =0.0029) Taxanes: Kappa = 0.75 (95% CI, 0.42 to 1.00, p =0.0028) Anthracyclines: R^2 = 0.49, p =0.0019 $\log_{10}HR_{OS} = -0.011 + 0.259\log_{10}HR_{PFS}$ Taxanes: R^2 = 0.35, p =0.012 $\log_{10}HR_{OS} = 0.014 + 0.499\log_{10}HR_{PFS}$
Sherrill et al. 2008[29]	mBC	Systematic search	RCTs published after 1994	67 (17,081)	Treatment effects for TTP/PFS and OS (HR-1) Significance of treatment effect in TTP/PFS and OS	Linear regression (through origin) on treatment effect weighted by sample size Unweighted Pearson correlation between HR Kappa test for agreement on significant treatment effect	Slope = 0.32 (95% CI, 0.20 to 0.43), R^2 = 0.30 R = 0.46 Kappa = 0.47, p <0.05
Wilkerson and Fojo 2009[30]	mBC mCRC mOC	"non - exhaustive" search	Randomised trials showing a statistically significant difference in either PFS or OS or their HRs	66 (NS)	Differences in median PFS and OS HR for PFS and OS	Linear regression on differences in medians Linear regression HR_{PFS} vs HR_{OS}	Slope = 1.214 (95%CI 0.89 to 1.54), R^2 = 0.49, p <0.0001 mCRC: R^2 = 0.61 p < 0.0001 mOC: R^2 = 0.60 p = 0.0007 mBC: R^2 = 0.30 p = 0.018 R^2 = 0.62, p <0.0001 mCRC: R^2 = 0.52 p = 0.0021 mOC: R^2 = 0.73 p = 0.02 mBC: R^2 = 0.70 and p = 0.0015
Bowater et al. 2011[26]	mBC mCRC (also locally	Systematic search	RCTs published in English between 1998 and 2008	mBC: 95 (NS)	Gain (%) in median TTP and PPS (PPS = median OS –	Spearman's rank correlation for gain	mBC: ρ = 0.37 mCRC: ρ = 0.11

	advanced disease)		comparing two different chemotherapy treatments	mCRC: 74 (NS)	median TTP)	Hypothesis (sign) test for proportion of trials with a) PPS%gain < TTP%gain b) PPS%gain <0.5TTP%gain	a) p <0.01 for both tumour types b) p<0.01 for both tumour types
Hotta et al. 2011[3]	Advanced or metastatic NSCLC	Systematic search	Phase III trials of first-line therapy	70 (38,721)	Median OS, PFS and PPS (PPS= median OS – median PFS)	Linear regression analysis weighted by trial size	Median OS and PFS: R^2 = 0.2563 Median OS and PPS: R^2 = 0.8917
Chirila et al. 2012[4]	mCRC	Systematic search	Randomised phase II and III trials with at least 20 participants	62 (23,527)	Median PFS/TTP and OS HR for PFS/TTP and OS (HR defined as ratio of medians)	Pearson product-moment correlation Spearman's rank correlation Weighted least squares regression weighted by trial size Diagnostic evaluation of regression equations (ROC curves for outcome of HRos ≤0.8)	PFS: 0.89 (95%CI 0.83 – 0.93) TTP: 0.75 (95%CI 0.59 – 0.84) PFS/TTP: 0.87 (95%CI, 0.82 to 0.91) PFS: 0.78 (95%CI, 0.66 to 0.85) TTP: 0.59 (95%CI 0.37 – 0.74) PFS/TTP: 0.76 (95%CI, 0.67 to 0.82) Ratio of Medians PFS/TTP and OS: Slope = 0.41 (95%CI, 0.30 to 0.52), intercept = 0.60 (95%CI, 0.49 to 0.71), R^2 = 0.48 Ratio of Medians PFS and OS: Slope = 0.49 (95%CI, 0.35 to 0.64), intercept = 0.52 (95%CI, 0.39 to 0.66), R^2 = 0.59 Ratio of Medians TTP and OS: Slope = 0.31 (95%CI, 0.12 to 0.49), intercept = 0.71 (95%CI, 0.53 to 0.90), R^2 = 0.32 AUC = 0.795 (p <0.01) HR_{PFS} ≤0.78 has sensitivity =0.89 and specificity=0.69
Shitara et al. 2012[5]	Advanced gastric	Systematic search	Randomised phase II and III trials of systemic chemotherapy	36 (10,484)	Median PFS/TTP and OS HR of PFS/TTP and OS	Spearman's rank correlation (also by subgroups)	Median PFS/TTP and OS: ρ = 0.70 (95%CI, 0.59 to 0.82), p <0.001 HR PFS/TTP and OS: ρ = 0.80 (95%CI, 0.68 to 0.92), p <0.0001
Sundar et al.	mOC	Systematic	Any randomised controlled	37	Gain (%) in median PFS/TTP	Spearman's rank correlation for	Gain in median PFS/TTP and PPS

2012[31]		search	trials of chemotherapy in treating metastatic ovarian cancer	(15,850)	and PPS	gain Hypothesis (sign) test for proportion of trials with a) PPS%gain = 0 b) PPS%gain > PFS/TTP %gain	in primary treatment: ρ = 0.06, p = 0.69 Gain in median PFS/TTP and PPS at recurrence: ρ = −0.234 (95%CI, -0.73 to 0.43), p = 0.49 a) p =0.85 in primary treatment, p = 0.99 at recurrence b) p = 0.23 at recurrence
Amir et al. 2012[32]	mPancreatic mNSCLC, mCRC, mRCCl, mHNC mBC mOC	Purposive sampling of RCTs	RCTs supporting registration of new anti-cancer drugs approved by the US FDA in the last 10 years	26 (NS)	HR of PFS/TTP and OS	Linear regression weighted by the trial sample size (Pearson coefficient)	HR for OS and PFS: R = 0.64 for the group with PPS<12 months R = 0.38 for the group with PPS≥12 months
Li et al. 2012[b]	Advanced NSCLC	Systematic search	Phase II and Phase III (randomised and non randomised) Clinical trials published before August 2011 assessing gefitinib or erlotinib monotherapy	60 (9,903)	Median PFS or TTP and Median Survival time	Linear regression weighted by the trial sample size, also adjusted by covariates ROC analysis (AUC) to examine accuracy in prediction of MST	PFS and MST: R^2 = 0.70, p < 0.0001 R^2 = 0.74, p < 0.001 (adjusted) PFS and MST (adjusted): R^2 = 0.89, p < 0.001 Slope = 1.74, s.e. = 0.25 TTP and MST: R^2 = 0.04, p = 0.512 AUC_{PFS} = 81.5, p = 0.076 AUC_{PFS} = 94, p = 0.842 (adjusted)
Hayashi et al. 2012[7]	Advanced or mNSCLC	Systematic search	RCTs phase III published in English between 2000 and April 2011 that compared two or more systemic chemotherapies in patients with disease recurrence after chemotherapy	18 (11,310)	Median OS, median PFS/TTP, median PPS Incremental gains in median OS and median PFS/TTP	Spearman's rank correlation (weighted by the number of patients in each arm)	Median PFS/TTP and median OS: ρ = 0.51, p = 0.001 Absolute gains in median OS and median PFS/TTP: ρ = 0.29, p < 0.0001
Delea et al. 2012[33]	mRCC	Systematic search	Clinical trials published in English between 1997 and 2010	31 (10,943)	Absolute differences between median PFS/TTP, PFS or TTP and median OS Negative of the LogHR for PFS/TTP, PFS or TTP and OS	Pearson correlation coefficients (Multivariate) Ordinary least squares regression (weighted by samples size or inverse of the variance)	Absolute difference in median PFS/TTP and OS: ρ = 0.54, p = 0.0002 Intercept = 0.13 (95% CI, -1.44 to 0.77) Slope = 1.17 (95%CI, 0.59 to 1.76) R^2 = 0.28 PFS and OS: ρ = 0.55, Slope = 1.21 (95%CI, 0.56 to 1.86) R^2 = 0.28 TTP and OS: ρ = -0.10, Slope = -

							0.21 (95%CI, -2.98 to 2.56) R^2 = -0.24 –log$HR_{PFS/TTP}$ and -logHR_{OS}: ρ = 0.80, p < 0.0001 Intercept = -0.04 (95% CI, -0.12 to 0.04) Slope = 0.64 (95%CI, 0.08 to 0.47) R^2 = 0.63 -logHR_{PFS} and -logHR_{OS}: ρ = 0.81, Slope = 0.68 (95%CI, 0.49 to 0.86) R^2 = 0.65 -logHR_{TTP} and -logHR_{OS}: ρ = 0.64, Slope = 0.17 (95%CI, -0.20 to 0.53) R^2 = 0.21
Individual patient level data							
Buyse et al. 2007[8]	Advanced CRC	Not stated but all had individual patient data	RCTs with a FU+leucovorin treatment arm	Historic: 10 (3,089) Validation: 3 (1,263)	Individual level: 6 months PFS and 12 months OS PFS and OS over entire time range Trial level: HR for PFS and OS	Rank correlation coefficient for PFS at 6 months and OS at 12 months Rank correlation coefficient for PFS and OS for entire time range Linear regression for treatment effects (logHR) on PFS and OS STE	ρ = 0.32 (95% CI,-0.14 to 0.67) ρ = 0.82 (95% CI, 0.82 to 0.83) R was equal to 0.99 (95% CI, 0.94 to 1.04) (R^2 = 0.98) log HR_{OS} = 0.003 + 0.81xlog HR_{PFS} STE HR_{PFS} = 0.86
Burzykowski et al. 2008[9]	mBC	Not stated but all had individual patient data	Randomised trials comparing anthracycline with taxane (both single agent and combination therapy)	11 (3,953)	Individual level: PFS, TTP and OS Trial level: HR for PFS, TTP and OS	Spearman's rank correlation coefficient for correlation between endpoints Spearman's rank correlation coefficient for treatment effects (HR) on endpoints Hougaard copula model of the relationship between treatment effects (logHR)	Individual PFS and OS: ρ = 0.688; (95% CI, 0.686 to 0.690) Individual TTP and OS: ρ = 0.682; (95% CI, 0.680 to 0.684) LogHR for PFS and OS: ρ = 0.48 (95% CI, -0.34 to 1.30) LogHR for TTP and OS: ρ = 0.49 (95% CI,-0.32 to 1.30) Regression parameters not reported
Foster et al. 2011[17]	SCLC	Consecutive trials from the NCCTG	First-line trials (phase II and III), randomised and non randomised, that included either a platinum or taxol based regimen	9 (870)	Individual level: PFS status at 2,4,6 months and OS Trial level: LogHR by trial centre (32	Individual: Multivariate landmark analysis for OS by PFS at 2,4,6 months and c-index Trial level:	Individual: 2month: HR = 0.40 (95%CI, 0.30 to 0.52), c-index = 0.60 4month: HR = 0.42 (95%CI, 0.35 to 0.51), c-index = 0.63 6month: HR = 0.41 (95%CI, 0.35 to

					units) for PFS and OS	Weighted least square regression Spearman correlation coefficient Bivariate survival model (Copula)	0.49), c-index = 0.65 Trial level: WLS R^2 = 0.79 Spearman ρ = 0.75 Copula R^2 = 0.80
Halabi et al. 2009[18]	Progressive castrate-resistant prostate Cancer	Not stated	Phase II and III multicentre trials conducted by CALGB	9 (1,296)	Individual patient data on PFS and OS	Landmark analysis for OS by PFS at 3 months, 6 months Kendall τ for association between PFS and OS	3month PFS: HR = 2.0 (95% CI, 1.7 to 2.4; p <0.001) 6month PFS: 1.9 (95% CI, 1.6 to 2.4; p <0.001) τ = 0.30 (bootstrap s.e. = 0.0172, 95% CI, 0.26 to 0.32, p <0.00001)
Heng et al. 2011[19]	mRCC	Not relevant	Consecutive population based samples treated on clinical trial or off protocol at 12 cancer centres	NS (1,158)	Individual patient data on PFS and OS	Landmark analysis of OS by PFS at 3 months, 6 months Kendall τ for PFS and OS Fleischer's model correlation	3month: HR = 3.05 (95% CI, 2.42 to 3.84) 6month: HR = 2.96 (95% CI, 2.39 to 3.67) 0.42 (bootstrap s.e., 0.016, 95% CI, 0.39 to 0.45, p <.0001) 0.66 (bootstrap s.e., 0.025, 95% CI, 0.61 to 0.71)
Polley et al. 2010[20]	Brain (GBM)	Not relevant	Phase II trials conducted at a single institution	3 (193)	Individual patient data on PFS and OS	Landmark analysis for OS by PFS at 10 weeks, 18 weeks, 26 weeks	10weeks: HR = 3.55 (95%CI, 2.28 to 5.52) 18weeks: HR = 2.06 (95%CI, 1.43 to 2.99) 26weeks: HR = 1.99 (95%CI, 1.38 to 2.85) (combined across all trials)
Mandrekar et al. 2010[21]	Advanced NSCLC	Not relevant	Consecutive NCCTG phase II trials	4 (284)	Individual patient data on PFS and OS	Landmark analysis for OS by PFS at 8 weeks, 12 weeks, 16 weeks, 20 weeks, 24 weeks	8 weeks: HR = 0.45 (95%CI, 0.33 to 0.62), p <0.0001, c-index = 0.63 12 weeks: HR = 0.39 (95%CI, 0.28 to 0.52), p <0.0001, c-index = 0.67 16 weeks: HR = 0.49 (95%CI, 0.36 to 0.65), p <0.0001, c-index = 0.66 20 weeks: HR = 0.41 (95%CI, 0.30 to 0.55), p <0.0001, c-index = 0.68 24 weeks: HR = 0.41 (95%CI, 0.30 to 0.57), p <0.0001, c-index = 0.68
Green et al.	Advanced	Not stated but all	NS	10	Rate of $PFS_{1\text{-year}}$ and $OS_{2\text{-year}}$,	Per-patient agreement between	$PFS_{1\text{-year}}$ and $OS_{2\text{-year}}$:

2008[10]	CRC	had individual patient data		(NS)	$OS_{5\text{-year}}$ HR of $PFS_{1\text{-year}}$ and $OS_{2\text{-year}}$	endpoints (%) Study-wise agreement Linear regression weighted by the trial sample size Spearman's rank correlation Individual-level correlation estimated using a bivariate survival model Trial-level correlation estimated using a bivariate survival model Proportion of treatment effect (PTE) on OS explained by PFS	Agreement = 89% 8/10 trials yield same conclusions R^2 = 0.002 $OS_{2\text{-year}}$ rate= 0.21 + 0.03 x $PFS_{1\text{-year}}$ rate Slope s.e. = 0.19, p >0.20; Intercept s.e. = 0.03, p <0.001 ρ = 0.13 $HR_{PFS1\text{-year}}$ and $HR_{OS2\text{-year}}$: R^2 = 0.84 $HR_{OS2\text{-year}}$ = 0.44 + 0.57 x $HR_{PFS1\text{-year}}$ Slope s.e. = 0.09, p = 0.0002; Intercept s.e. = 0.122, p =0.007 ρ = 0.92 $HR_{PFS1\text{-year}}$ and $HR_{OS2\text{-year}}$: R^2_{indiv} = 0.61 (95% CI, 0.59 to 0.64) R^2_{trial} = 0.58 (95% CI, 0.18 to 0.98) PTE > 100%
Burzykowski and Buyse 2006[35]	Advanced CRC Advanced ovarian	Not stated but all had individual patient data (same as Burzykowski et al. 2001)	NS	CRC: 2 (642) OC: 4 (1,194)	CRC: Center-based HR of PFS and OS (log scale) OC: Center-based for the two larger trials, and trial-based for the two smaller trials HR of PFS and OS (log scale)	Hougaard copula model of the relationship between treatment effects (log scale) Surrogate threshold effect (using estimates for model parameters and prediction variance to correct for estimation)	Advanced colorectal: $LogHR_{PFS}$ = 0.021, Var = 1.149 $LogHR_{OS}$ = 0.003, Var = 0.737 R^2_{Trial} = 0.53 (95% CI, 0.34 to 0.72) R^2_{Trial} = 0.64 (adjusted for the estimation error in treatment effects) Advanced ovarian: $LogHR_{PFS}$ = -0.20, Var = 1.02 $LogHR_{OS}$ = -0.18, Var = 0.93 R^2_{Trial} = 0.88 (95% CI, 0.81 to 0.95) R^2_{Trial} = 0.83 (adjusted for the estimation error in treatment effects) Advanced colorectal: STE on $logHR_{PFS}$ = -2.11 STE on $logHR_{PFS}$ = -3.11 (adjusted) STE on HR_{PFS} = 0.12 STE on HR_{PFS} = 0.04 (adjusted)

							Advanced ovarian: STE on $logHR_{PFS}$ = -0.75 STE on $logHR_{PFS}$ = -0.61 (adjusted) STE on HR_{PFS} = 0.47 STE on HR_{PFS} = 0.54 (adjusted)
Ballman et al. 2007[11]	Brain (GBM)	All trials of newly diagnosed and recurrent GBM conducted by the NCCTG	Trials conducted by the NCCTG on newly diagnosed and recurrent GBM patients	27 (1,693) Newly diagnosed: 11 (1,348) Recurrent: 16 (345)	$PFS_{6months}$ and $OS_{12months}$ $PFS_{6months}$ and OS	Patient-level agreement Kappa statistics Linear regression weighted by the trial sample size Landmark analysis OS by $PFS_{6months}$	Newly diagnosed GBM: 75% K = 0.48 (95% CI, 0.44 to 0.53) Recurrent GBM: 88% K = 0.52 (95% CI, 0.39 to 0.65) Newly diagnosed GBM: $OS_{12months} = 0.24 + 0.40 \times PFS_{6months}$ p slope = 0.09 $R^2 = 0.28$ Recurrent GBM: $OS_{12months} = 0.08 + 0.61 \times PFS_{6months}$ p slope = 0.01 $R^2 = 0.41$ Newly diagnosed: HR = 2.1 (95% CI, 1.8 to 2.4) Recurrent: HR = 2.4 (95% CI, 1.6 to 3.8)
Rose et al. 2010[13]	mOC	Exploratory data analysis	A series of consecutive GOG second-line phase II trials in the setting of platinum-resistant cancer	11 (407)	Aggregate $PFS_{6months}$ rates and median OS	Pearson correlation coefficient Kendall τ-b correlation coefficient	$PFS_{6months}$ and median OS: Pearson r = 0.661, p = 0.027 Kendall τ-b r = 0.514, p = 0.029
Buyse et al. 2000[14]	Advanced ovarian	Trials in the Ovarian Cancer Meta-analysis Project. All had individual patient data.	Not stated	4 (1,194)	Individual level: LogTTP and LogOS Trial level: TE on LogTTP and LogOS (absolute difference)	Prentice criteria tests of significance of association between endpoints Freedman's proportion explained Relative effect Adjusted association Random effects meta-analytic model of jointly normally distributed endpoints	LogTTP and LogOS: α, p = 0.003; β, p = 0.054; γ, p < 0.0001 PE = 1.46 (95% CI, 0.80 to 2.13) RE = 0.60 (95% CI, 0.32 to 0.87) ρ_Z = 0.942 (95% CI, 0.94 to 0.95) R^2_{trial} = 0.951, s.e. = 0.098 R^2_{indiv} = 0.888, s.e. = 0.006
Burzykowski et al. 2001[12]	Advanced CRC Advanced ovarian	OC: Trials in the Ovarian Cancer Meta-analysis Project. All had individual patient	Not stated	CRC: 2 (642) OC: 4 (1,153)	Individual level: PFS and OS Trial level: CRC: Center-based HR of	Clayton's copula model for the association between two failure time endpoints with common base-line hazard	Advanced ovarian: R^2_{Trial} = 0.95 (95% CI, 0.76 to 1.14) (adjusted for the estimation error in treatment effects) τ = 0.857 (95%CI, 0.845 to 0.870)

		data			PFS and OS OC: Center-based for the two larger trials, and trial-based for the two smaller trials HR of PFS and OS	Hougaard's copula model for the association between two failure time endpoints with common base-line hazard	Advanced colorectal: R^2_{Trial} = 0.24 (95% CI, -0.40 to 0.89) (adjusted for the estimation error in treatment effects) τ = 0.502 (95%CI, 0.457 to 0.548) Advanced ovarian: R^2_{Trial} = 0.95 (95% CI, 0.82 to 1.07) (adjusted for the estimation error in treatment effects) τ = 0.839 (95%CI, 0.828 to 0.850) Advanced colorectal: R^2_{Trial} = 0.33 (95% CI, -0.69 to 1.36) (adjusted for the estimation error in treatment effects) τ = 0.583 (95%CI, 0.548 to 0.619)

AUC = area under the curve; CALGB = Cancer and Leukemia Group B; mBC = metastatic breast cancer; mCRC = metastatic colorectal cancer; FAC = 5-fluorouracil, adriamycin and cyclophosphamide; FEC = 5-fluorouracil, epirubicin and cyclophosphamide; FU = fluorouracil; GBM = Glioblastoma multiforme; GOG = Gynecologic Oncology Group; HR = hazard ratio; HRP = hormone refractory prostate; MST = Median Survival time; NCCTG = North Central Cancer Treatment Group; NS = not stated; mNSCLC = metastatic Non-small cell lung cancer; mOC = metastatic ovarian cancer; OS = overall survival; PFS = progression free survival; PPS = post-progression survival; mRCC = metastatic renal cell carcinoma; RCT=randomised controlled trial; ROC = receiver operating characteristic; SCLC = small cell lung cancer; s.e. = standard error; STE = Surrogate Threshold Effect; TE = Treatment effect, TTP = time to progression; WLS = weighted list squares; Var = variance.

Supplementary Table 3 Factors considered in multivariate analyses

Reference	Factors analysed	
Johnson et al. 2006[24]	▪ Patients' age (median) ▪ Performance status ▪ Stage of disease	▪ Year of trial ▪ Trial methodological quality ▪ Use of rescue (or salvage) treatment
Chirila et al. 2012[4]	▪ Line of therapy ▪ Performance status ▪ Clinical trial phase ▪ Crossover after progression	▪ Drug therapy ▪ Publication year ▪ Median OS for the control group
Hackshaw et al.[23]	▪ Before/after 1990 when second line therapies not commonly used	▪ Death included in surrogate time-to-event outcome (i.e PFS not TTP)
Sherrill et al. 2008[29]	▪ Treatment class (hormonal, anthracyclines, first-line, non-first-line) ▪ Only HER2+ patients ▪ Study size (>100 per arm) ▪ TTP >6 mths in control arm	▪ Reported HRs ▪ ITT analyses ▪ Blinding
Miksad et al. 2008[28]	▪ Strict PFS definition ▪ Year last patient recruited	▪ First / subsequent line treatment
Hotta et al. 2009[27]	▪ Year of trial ▪ Old agents used ▪ Cisplatin used ▪ Carboplation used ▪ Full publication or abstract ▪ Description of sample size calculation	▪ Definition of primary endpoint ▪ Description of TTP definition ▪ Description of OS definition ▪ Description of definition for both TTP and OS ▪ Sample size
Shitara et al. 2012[5]	▪ PFS or TTP ▪ Trial area (Asian or non-Asian) ▪ Before 2006 or after 2006 ▪ <200 or ≥200 patients	▪ Registration trial with investigational agents ▪ Number of chemotherapeutic agents in treatment arm ▪ Proportion of measurable disease ▪ Proportion of patients who received second-line chemotherapy
Li et al. 2012[6]	▪ Lines of therapy ▪ Patients origin ▪ Proportions of female patients	▪ Never-smokers ▪ Patients with adenocarcinoma histology ▪ Patients with performance status ≥ 2
Delea et al. 2012[33]	▪ Prior treatment ▪ Targeted therapy ▪ TTP or PFS ▪ Crossover allowed	▪ Year of publication ▪ <200 or ≥200 patients ▪ HR estimated from Kaplan-Meier curves ▪ Drug class

Supplementary references

1. Louvet C, de Gramont A, Tournigand C, Artru P, Maindrault-Goebel F, Krulik M. Correlation between progression free survival and response rate in patients with metastatic colorectal carcinoma. *Cancer* 2001;91:2033-8.

2. Tang PA, Bentzen SM, Chen EX, Siu LL. Surrogate end points for median overall survival in metastatic colorectal cancer: literature-based analysis from 39 randomized controlled trials of first-line chemotherapy. *Journal of clinical oncology* : official journal of the American Society of Clinical Oncology 2007;25:4562-8.

3. Hotta K, Kiura K, Fujiwara Y, et al. Role of survival post-progression in phase III trials of systemic chemotherapy in advanced non-small-cell lung cancer: a systematic review. *PloS one* 2011;6:e26646.

4. Chirila C, Odom D, Devercelli G, et al. Meta-analysis of the association between progression-free survival and overall survival in metastatic colorectal cancer. *International journal of colorectal disease* 2012;27:623-34.

5. Shitara K, Ikeda J, Yokota T, et al. Progression-free survival and time to progression as surrogate markers of overall survival in patients with advanced gastric cancer: analysis of 36 randomized trials. *Investigational New Drugs* 2012;30:1224-31.

6. Li X, Liu S, Gu H, Wang D. Surrogate end points for survival in the target treatment of advanced non-small-cell lung cancer with gefitinib or erlotinib. *Journal of Cancer Research & Clinical Oncology* 2012; 138:1963-9.

7. Hayashi H, Okamoto I, Taguri M, Morita S, Nakagawa K. Postprogression survival in patients with advanced non-small-cell lung cancer who receive second-line or third-line chemotherapy. *Clinical lung cancer* 2013;14:261-6.

8. Buyse M, Burzykowski T, Carroll K, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of clinical oncology* : official journal of the American Society of Clinical Oncology 2007;25:5218-24.

9. Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. *Journal of clinical oncology* : official journal of the American Society of Clinical Oncology 2008;26:1987-92.

10. Green E, Yothers G, Sargent DJ. Surrogate endpoint validation: statistical elegance versus clinical relevance. *Statistical Methods in Medical Research* 2008;17:477-86.

11. Ballman KV, Buckner JC, Brown PD, et al. The relationship between six-month progression-free survival and 12-month overall survival end points for phase II trials in patients with glioblastoma multiforme. *Neuro-oncology* 2007;9:29-38.

12. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society*: Series C (Applied Statistics) 2001;50:405-22.

13. Rose PG, Tian C, Bookman MA. Assessment of tumor response as a surrogate endpoint of survival in recurrent/platinum-resistant ovarian carcinoma: a Gynecologic Oncology Group study. *Gynecologic Oncology* 2010;117:324-9.

14. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000;1:49-67.

15. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. Use of meta-analysis for the validation of surrogate endpoints and biomarkers in cancer trials. *Cancer Journal* 2000;15:421-5.

16. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. *Journal of Clinical Oncology* 2008;26:3913-5.

17. Foster NR, Qi Y, Shi Q, et al. Tumor response and progression-free survival as potential surrogate endpoints for overall survival in extensive stage small-cell lung cancer: findings on the basis of North Central Cancer Treatment Group trials. *Cancer* 2011;117:1262-71.

18. Halabi S, Vogelzang NJ, Ou SS, Owzar K, Archer L, Small EJ. Progression-free survival as a predictor of overall survival in men with castrate-resistant prostate cancer. *Journal of clinical oncology* : official journal of the American Society of Clinical Oncology 2009;27:2766-71.

19. Heng DY, Xie W, Bjarnason GA, et al. Progression-free survival as a predictor of overall survival in metastatic renal cell carcinoma treated with contemporary targeted therapy. *Cancer* 2011;117:2637-42.

20. Polley MY, Lamborn KR, Chang SM, Butowski N, Clarke JL, Prados M. Six-month progression-free survival as an alternative primary efficacy endpoint to overall survival in newly diagnosed glioblastoma patients receiving temozolomide. *Neuro-oncology* 2010;12:274-82.

21. Mandrekar SJ, Qi Y, Hillman SL, et al. Endpoints in phase II trials for advanced non-small cell lung cancer. Journal of thoracic oncology : official publication of the International Association for the Study of Lung *Cancer* 2010;5:3-9.

22. Fleischer F, Gaschler-Markefski B, Bluhmki E. A statistical model for the dependence between progression-free survival and overall survival. *Statistics in Medicine* 2009;28:2669-86.

23. Hackshaw A, Knight A, Barrett-Lee P, Leonard R. Surrogate markers and survival in women receiving first-line combination anthracycline chemotherapy for advanced breast cancer. *British journal of cancer* 2005;93:1215-21.

24. Johnson KR, Ringland C, Stokes BJ, et al. Response rate or time to progression as predictors of survival in trials of metastatic colorectal cancer or non-small-cell lung cancer: a meta-analysis. *The lancet oncology* 2006;7:741-6.

25. Bowater RJ, Bridge LJ, Lilford RJ. The relationship between progression-free and post-progression survival in treating four types of metastatic cancer. *Cancer letters* 2008;262:48-53.

26. Bowater RJ, Lilford PE, Lilford RJ. Estimating changes in overall survival using progression-free survival in metastatic breast and colorectal cancer. *International journal of technology assessment in health care* 2011;27:207-14.

27. Hotta K, Fujiwara Y, Matsuo K, et al. Time to progression as a surrogate marker for overall survival in patients with advanced non-small cell lung cancer. *Journal of thoracic oncology* : official publication of the International Association for the Study of Lung Cancer 2009;4:311-7.

28. Miksad RA, Zietemann V, Gothe R, et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. *International journal of technology assessment in health care* 2008;24:371-83.

29. Sherrill B, Amonkar M, Wu Y, et al. Relationship between effects on time-to-disease progression and overall survival in studies of metastatic breast cancer. *British journal of cancer* 2008;99:1572-8.

30. Wilkerson J, Fojo T. Progression-free survival is simply a measure of a drug's effect while administered and is not a surrogate for overall survival. *Cancer journal* 2009;15:379-85.

31. Sundar S, Wu J, Hillaby K, Yap J, Lilford R. A systematic review evaluating the relationship between progression free survival and post progression survival in advanced ovarian cancer. *Gynecologic Oncology* 2012;125:493-9.

32. Amir E, Seruga B, Kwong R, Tannock IF, Ocana A. Poor correlation between progression-free and overall survival in modern clinical trials: are composite endpoints the answer? European Journal of *Cancer* 2012;48:385-8.

33. Delea TE, Khuu A, Heng DYC, Haas T, Soulieres D. Association between treatment effects on disease progression end points and overall survival in clinical studies of patients with metastatic renal cell carcinoma. *British journal of cancer* 2012;107:1059-68.

34. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* (Clinical research ed) 1997;315:629-34.

35. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics* 2006;5:173-86.

36. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *International Journal of Clinical Oncology* 2009;14:102-11.

37. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine* 2006;25:183-203.