A REVIEW OF STUDIES EXAMINING THE RELATIONSHIP BETWEEN PROGRESSION-FREE SURVIVAL AND OVERALL SURVIVAL IN ADVANCED OR METASTATIC CANCER

REPORT BY THE DECISION SUPPORT UNIT

August 2012

Sarah Davis, Paul Tappenden, Anna Cantrell

School of Health and Related Research, University of Sheffield

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street
Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University. The DSU is commissioned by The National Institute for Health and Clinical Excellence (NICE) to provide a research and training resource to support the Institute's Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

The production of this document was funded by NICE through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

Acknowledgements

We thank Oriana Ciani, Bhash Naidoo and Allan Wailoo for their helpful peer review comments and Jenny Dunn for formatting the report. The content of this report is the responsibility of the authors alone.

EXECUTIVE SUMMARY

Introduction

Progression-free survival (PFS), time-to-progression (TTP) and overall survival (OS) are commonly used endpoints in randomised controlled trials (RCTs) and observational studies of treatments for metastatic disease in solid tumour cancers. Within health economic models of cancer interventions, disease- and treatment-related health states are often defined according to whether the patient has experienced disease progression since starting a particular line of therapy. Such models typically estimate the mean sojourn time in these health states and therefore information is needed on both PFS/TTP and OS. PFS or TTP are sometimes regarded as valid surrogate outcomes when establishing the clinical benefit of a treatment in the absence of mature data on OS, but an estimate of OS is still needed within the economic analysis. Some quantification of the relationship between PFS/TTP and OS may be used to populate the economic model as an alternative to directly modelling OS from the trial data.

Objectives

The aim of this review was to examine the evidence available concerning the relationship between PFS/TTP and OS in advanced or metastatic cancer, with a view to determining the extent to which PFS/TTP can be considered a robust surrogate endpoint for OS.

Methods

A review was conducted of papers using meta-regression or other meta-analytic techniques to examine the statistical relationship between OS and either PFS or TTP. Studies were identified through citation searching as a systematic search was not considered feasible. The review was not restricted to any particular cancer type but instead included any form of cancer in which the treatment intent is palliative rather than curative and therefore the surrogate outcome of interest is progression-free rather than disease-free survival. The main focus was therefore advanced or metastatic cancer.

Results

Nineteen papers were included in the review covering eight different tumour types. Four studies used aggregate data from multiple trials to examine the relationship between PFS and OS for individual trial arms. Seven studies used individual patient data to examine the relationship between PFS and OS for individual patients. Thirteen studies examined the trial level relationship between treatment

effect on PFS and treatment effect on OS with three using individual patient data, whilst the remainder used aggregate data from multiple trials. A variety of statistical techniques were employed within these studies, with the most commonly used being rank correlation coefficients, linear regression and landmark analysis. The lack of a standardised approach made it difficult to establish whether there is a consistent relationship between PFS/TTP and OS. The majority of the studies found a positive correlation between PFS/TTP and OS for individual patients, individual trial arms and the treatment effect between trial arms, although, the size of the correlation and its statistical significance varied considerably across studies. This is not surprising given the variety of methods employed and the variation in studies characteristics such as differences in the tumour type and the line of therapy considered.

Conclusion

This review suggests that the level of evidence available supporting a relationship between PFS/TTP and OS varies considerably by cancer type and is not always consistent even within one specific cancer type. Furthermore, even where strong consistent evidence supporting a correlation between the treatment effects (i.e Level 1 evidence according to Elston and Taylor's framework) is available, it is unclear how that should be converted into a quantified relationship between PFS and OS treatment effects within a cost-effectiveness model. Therefore, any cost-effectiveness analysis which makes a strong assumption regarding the relationship between PFS and OS should be treated with caution. We would support Elston and Taylor in recommending that any cost-effectiveness analysis based on a surrogate relationship between PFS and OS should be supported with a transparent explanation of how the relationship is quantified in the model and should be accompanied by sensitivity analysis exploring the uncertainty associated with that relationship and a systematic review of papers examining the relationship between PFS and OS in the relevant setting. This would allow decision makers to judge the appropriateness of the model in light of the evidence available in that specific disease area.

CONTENTS

TABLES AND FIGURES

1. INTRODUCTION

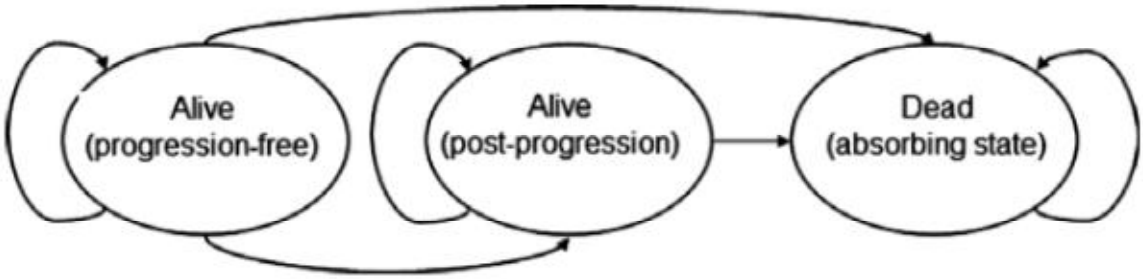
1.1. BACKGROUND

Progression-free survival (PFS) and overall survival (OS) are commonly used endpoints in randomised controlled trials (RCTs) and observational studies of treatments for metastatic disease in solid tumour cancers (e.g. breast, lung, prostate, colorectal). Within an RCT setting, OS is measured as the time from randomisation to death due to any cause and PFS is measured as the time from randomisation to either documented disease progression or death. Disease progression in solid tumours may involve an increase in the size of existing lesions or the appearance of new lesions.[1] OS is unambiguous as the true time of death can be measured accurately. PFS is associated with more complex issues as disease progression is typically measured relative to disease status at the start of treatment; where treatment involves more than one line of therapy, the time from treatment initiation to further progression may be considered the "progression-free" period for that line of therapy. Some clinical study publications report time-to-progression (TTP) rather than or alongside PFS, the difference being that in TTP, patients are censored at the point of death. In patients who have documented progression prior to death, post-progression survival (PPS) is defined as the time from progression to death. It therefore provides an indirect measure of OS in patients with a known TTP.

1.2. RELEVANCE OF PFS/TTP AND OS OUTCOMES TO COST-EFFECTIVENESS ANALYSIS

Both OS and PFS/TTP are time-to-event outcomes and are usually represented in publications and trial reports using Kaplan Meier survival curves. Health economic models often use this information as a means of defining disease- and treatment-related health states. Commonly, these differentiate between a progression-free period and a post-progression period. Figure 1 shows a simple Markov-type structure with separate health states for patients before and after progression. This definition of health states is often also useful for discriminating between different treatment periods – when a patient's tumour progresses they may move on to receive another line of active treatment or alternatively they may receive best supportive care. Therefore, treatment costs are likely to change when moving from one health state to another following disease progression. Health states may also differ in terms of health-related quality of life, meaning that the quality adjusted life years (QALYs) gained associated with spending a fixed time in any one state will vary.

Figure 1: Example of health states in a simple Markov model



1.3. QUANTIFYING THE RELATIONSHIP BETWEEN PFS AND OS IN COST-EFFECTIVENESS MODELS

Within cost-effectiveness analysis, mathematical modelling is used to estimate the mean sojourn time in these health states and therefore requires some quantification of the relationship between PFS/TTP and OS. However, this relationship is often unclear from available trial data for a number of reasons:

1. Patients may receive more than one line of therapy and the survival benefits attributable to an individual treatment are not directly measurable and may not be known at all.

2. PFS usually requires radiological assessment according to the study follow-up schedule. This means that the exact time of progression is unknown and most of the events are recorded at one of the scheduled follow-up points leading to "lumpy" interval-censored PFS curves. Conversely, the point of death can be measured more accurately.

3. Clinical trials are often statistically powered according to PFS which means that there may be few data points in the Kaplan Meier survival curve for OS leading to greater uncertainty in the shape of the survival curve.

4. PFS includes death as an event within the Kaplan Meier survival analysis thereby leading to correlation in the outcomes and some degree of double-counting if the events are used independently in the model.

5. For interim analyses of trials, or analyses of trials which have terminated prematurely, the level of censoring in both PFS and OS may be considerable.

6. An individual-level covariance matrix relating PFS to OS cannot be constructed without patient level data from trials which are usually not within the public domain and may therefore not be available to inform technology appraisals.

There are several options for representing this relationship within a health economic model. Firstly, one could assume that an incremental benefit in PFS for Treatment A vs Treatment B leads to an

equivalent incremental benefit in OS for Treatment A vs Treatment B (i.e. $\Delta PFS=\Delta OS$). This approach assumes that both the probability and duration of post-progression survival benefits are exactly equivalent between treatment groups. This is unlikely to hold as progression and death are competing events. An example of this approach can be seen in the manufacturer models developed to inform the appraisals of bevacizumab for the first-line treatment of metastatic colorectal cancer and rituximab for the first-line treatment of chronic lymphocytic leukaemia.[2,3]

Secondly one could assume that an incremental benefit in PFS for Treatment A vs Treatment B leads to a proportional gain in OS for Treatment A vs Treatment B (i.e. $\Delta PFS=\Delta OS.\alpha$). This is essentially an arbitrary judgement but may reflect a technical value judgment of the Appraisal Committee. This approach is evident in the manufacturer's submission within the appraisal of rituximab for the first-line treatment of Stage III-IV follicular non-Hodgkin's lymphoma.[4] Evidence on the correlation between treatment effects for median PFS and median OS, from a suitable set of trials, could be used to validate the relationship between these outcomes predicted by the economic model.

A third option, which avoids making explicit assumptions regarding the relationship between the endpoints, would involve modelling independent curves for PFS and OS for Treatment A and Treatment B. An example of this can be seen in the Assessment Group model of bevacizumab for the first-line treatment of metastatic colorectal cancer.[2] However, this approach may lead to problems, such as intersecting PFS and OS curves, if the data on OS are too immature to support a robust extrapolation.

1.4. EVALUATING THE RELATIONSHIP BETWEEN PFS/TTP AND OS

If PFS is to be used as a surrogate for OS in cost-effectiveness analyses, it is important to establish whether there is evidence of a relationship between PFS and OS that would support its suitability as a surrogate. Elston and Taylor recommend that evidence on the relationship between the surrogate and the final patient-related outcome should be systematically identified and presented according to the following hierarchy:[5]

Level 1: evidence demonstrating treatment effects on the surrogate correspond to effects on the patient-related outcome (from clinical trials);

Level 2: evidence demonstrating a consistent association between surrogate outcome and final patient-related outcome (from epidemiological/observational studies);

Level 3: evidence of biological plausibility of relationship between surrogate and final patient-related outcome (from pathophysiologic studies and/or understanding of the disease process).

They recommend that when there is Level 1 or 2 validation evidence, there may be consideration given to undertaking a cost-effectiveness modelling analysis based on a surrogate outcome. Elston and Taylor also make reference to two alternative evidence hierarchies for surrogate outcomes in which a distinction is made between surrogates which have evidence from trials in either the same or a different drug class. In the JAMA guide for surrogate outcomes,[6] a higher level of evidence is awarded where there is evidence from within the same drug class as opposed to other drug classes, and in the Outcome Measures in Rheumatology Clinical Trials (OMERACT) criteria[7] a higher level is awarded where there is evidence across multiple drug classes. Within the context of HTA, where one is concerned with whether the treatment effect on a surrogate outcome may be used within an economic model to predict the treatment effect on a final patient-related outcome, its seems reasonable to prefer evidence from within the same drug class over that from other drug classes. This view is supported by Fleming who cautions against translating conclusions regarding the validity of a surrogate from one disease area or drug class to another.[8]

Aims and objectives

The aim of this review was to examine the evidence available concerning the relationship between PFS or TTP and OS in advanced or metastatic cancer, with a view to determining the extent to which PFS/TTP can be considered a robust surrogate endpoint for OS in some types of solid tumours. This review has not been restricted to a single cancer area, but rather includes evidence in metastatic or advanced colorectal, breast, prostate, lung, and renal cancer and aggressive primary brain tumours. Studies which considered early stage cancer were excluded as the treatment intent is usually curative and therefore the surrogate outcome of interest is disease-free survival rather than progression free survival. The review focusses on the assessment of the relationship between the putative surrogate endpoints and the final endpoint at the trial-level, thus it considers studies which use meta-regression or other meta-analytic techniques to examine the statistical relationship between OS and either PFS or TTP, rather than individual studies which simply report both outcomes.

2. REVIEW METHODS

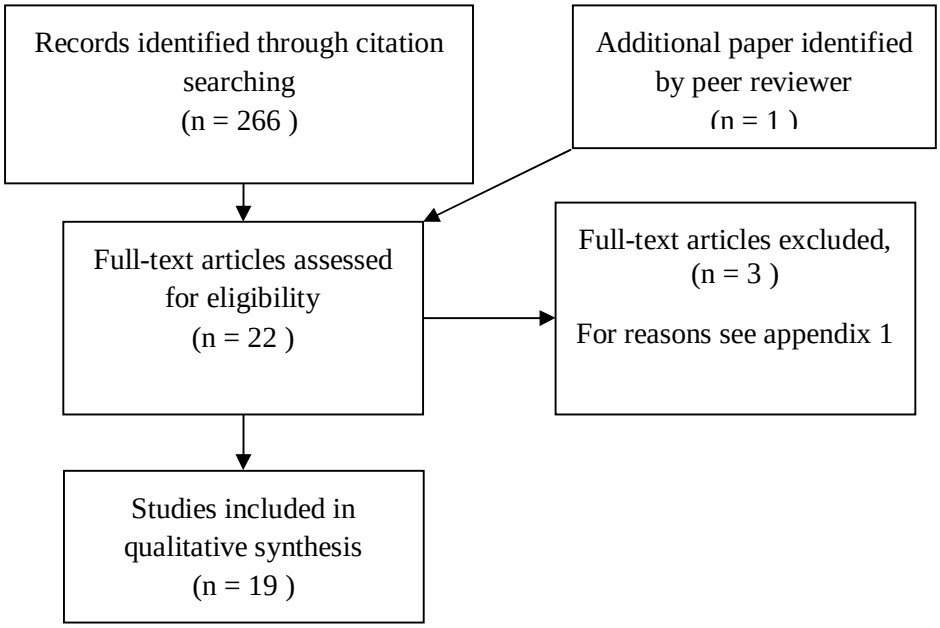
2.1. SEARCH STRATEGY

Citation searching was used to identify relevant papers from an initial list of three papers already known to the authors.[9-11] The citation searches were conducted on Medline and the Science Citation Index. The total number of papers sifted was 266. We explored the possibility of using a conventional systematic database search, but found that an exploratory search on Medline returned a very large number (over 3,000) of references. Attempts to make the search more specific resulted in many of the initially identified relevant papers and the papers identified through the citation search being excluded.

2.2. INCLUSION CRITERIA

Initially, the review was focused on metastatic disease in solid tumours, as PFS and OS are important trial endpoints in patients with metastatic disease where the treatment intent is palliative rather than curative. However, as "metastatic disease" is not a relevant term in all types of cancer (e.g Glioblastoma multiforme), we then broadened the review to include any form of cancer in which the treatment intent is palliative rather than curative and therefore the surrogate outcome of interest is progression-free rather than disease free survival. Papers were included in the review if they examined the statistical relationship between OS and either PFS or TTP. Papers simply reporting the target outcomes from single trials or multiple trials (e.g a meta-analysis for a particular drug or treatment class) were not included.[12] Conference abstracts were excluded from the review as they typically include limited information on the methods used.

Figure 2: Identification of included studies



3. STUDY CHARACTERISTICS

Nineteen papers were included in the review and these are summarised in Table 1. Eighteen papers were identified through the citation search (see Figure 2) and one additional recently published paper was identified by our peer reviewer.[13] Excluded studies and the reasons for exclusion can be found in Appendix 1. Twelve papers report analyses of aggregate trial outcomes from multiple RCTs. Six papers report analyses of individual patient data (IPD) from multiple trials, with two using exclusively Phase III trials,[14,15] two using a mixture of Phase II and Phase III trials[16,17] and two using exclusively Phase II trials.[18,19] A further paper was identified which analyses IPD from a retrospective cohort of patients not all of whom were enrolled in a trial.[20]

For the twelve papers using aggregate data from multiple trials, the number of trials included within the analysis varied from 13 to 191. For the papers using IPD, the number of included trials was lower and varied from 3 to 13. Most of the papers which used aggregate data from RCTs, reported using a systematic literature search to identify the trials although the authors of one study stated that their search was not considered to be exhaustive[21] and one study did not state how the included trials were identified.[22] The criteria used to select included trials varied markedly between papers with some

focusing on particular treatments[23,24] and others requiring only that treatment should be first-line.[9,22,25-27] Some restricted the analysis to trials of a certain size,[13,22,27] trials with "mature" survival data,[27] or trials which showed a statistically significant difference between treatment arms.[21]

The majority of the studies restricted their analysis to a single tumour type although some reported separate analyses for more than one tumour type[9,10,21,28] and one paper also reported a single analysis across multiple types of solid tumour.[21] The majority of papers describe the disease state as either metastatic or advanced. Within two studies, trials were included if there was a mixture of locally advanced and metastatic disease[23,28] but not all studies were explicit about whether they included studies with some patients with locally advanced disease. One study of patients with brain tumours (glioblastoma multiforme) was included despite the fact that the tumour type is described as being aggressive rather than advanced or metastatic. However, the poor median survival in this patient population (14.6 months) suggests that PFS and OS would be important outcomes in this population and that the treatment intent is palliative rather than curative.

Table 1: Characteristics of included papers

Reference	Tumour type	Study identification	Inclusion criteria	Number of studies	Number of patients
Louvet 2001[22]	Metastatic colorectal	Not stated	Phase III studies of first line treatment reported between 1990 and 2000, >100 patients per study arm	29	13,498
Hackshaw 2005[24]	Metastatic breast	Systematic search (Medline 1966-2005)	Randomised trials comparing FAC or FEC with one or more first-line combination therapies	42	9163
Johnson 2006[9]	Metastatic colorectal, metastatic non-small-cell lung	Systematic search	RCTs of first-line treatment	Colorectal: 146 Lung: 191	Colorectal: 35,557 Lung: 44,125
Buyse 2007[14]	Advanced colorectal	Not stated but all had individual patient data	RCTs with a FU+leucovorin treatment arm.	Historical: 10 Validation: 3	Historical: 3089 Validation; 1,263
Tang 2007[27]	Metastatic colorectal	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
Bowater 2008[10]	Metastatic breast, colorectal, hormone refractory prostate and non-small cell lung	Systematic search for reviews of RCT published between 1990 and 2007	RCTs comparing two different chemotherapy treatments	33, 38, 23 and 13 respectively by tumour type	Not stated
Burzykowski 2008[15]	Metastatic breast	Not stated but all had individual patient data	Randomised trials comparing anthracycline with taxane (both single agent and combination therapy)	11	3,953
Hotta 2009[26]	Advanced or metastatic non-small cell lung	Systematic search	Phase III trials of first-line therapies	54	23,457
Miksad 2008[23]	Advanced breast (some locally advanced included)	Systematic search	Randomised controlled trials of anthracyclines and taxanes	31	4,323
Sherrill 2008[11]	Metastatic breast	Systematic search	RCTs published after 1994	67	17,081

Reference	Tumour type	Study identification	Inclusion criteria	Number of studies	Number of patients
Wilkerson 2009[21]	Metastatic solid tumours	"non-exhaustive" search	Randomised trials showing a stat sig difference in either PFS or OS or their HRs	66	Not stated
Foster 2010[17]	"Extensive" small-cell lung	Consecutive trials from the North Central Cancer Treatment Group	First-line trials (phase II and III) that included either a platinum or taxol based regimen	9	870
Bowater 2011[28]	Metastatic breast, colorectal (included if proportion had locally advanced disease)	Systematic literature search	RCTs published between 1998 and 2008 comparing two different chemotherapy treatments	95 and 74 by tumour type respectively	Not stated
Hotta 2011[25]	Advanced or metastatic non-small cell lung	Systematic search	Phase III trials of first-line therapy	70	38,721
Halabi 2009[16]	Metastatic prostate	Not stated	Phase II and III multicentre trials conducted by CALGB	9	1,296
Heng 2011[20]	Metastatic renal cell	Not relevant	Consecutive population based samples at 12 cancer centres	Not relevant	1,158
Polley 2010[18]	Brain (Glioblastoma multiforme)	Not relevant	Phase II trials conducted at a single institution	3	193
Mandrekar 2010[19]	Advanced non-small cell lung	Not relevant	Consecutive NCCTG phase II trials	4	284
Chirila 2011[13]	Metastatic colorectal cancer	Systematic search	Randomised phase II and III trials with at least 20 participants	62	23,527

FAC=5-fluorouracil, adriamycin and cyclophosphamide, FEC=5-fluorouracil, epirubicin and cyclophosphamide, RCT=randomised controlled trial, FU= fluorouracil, CALGB=Cancer and Leukemia Group B, NCCTG=North Central Cancer Treatment Group

Table 2: Methods used to analyse the relationship between PFS/ TTP and OS

Reference	Data type	Relationships analysed	Methods used	Results
Louvet 2001[22]	Summary data from trials	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 2005[24]	Summary data from trials	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+0.5082x$log_{10}\ HR_{OS}$ (P<0.001, R^2=56%, s.e.=0.0928).
Johnson 2006[9]	Summary data from trials	Gain in median TTP and gain in median OS	Linear regression weighted by trial size (multivariate analysis used to explore other potential predictive factors) Surrogate threshold effect for various trial sizes	Colorectal: R^2=0·33; p<0·0001 OS = –0·002 + 0·0961 x TTP Lung: R^2=0·19; p=0·0003 OS = 0·189 + 0·616 x TTP 3.3 mths in colorectal and 3.2 mths in lung cancer for trial of 250 patients
Buyse 2007[14]	Individual patient level data	Individual level: 6mth PFS and 12mth OS PFS and OS over entire time range Trial level: Hazard ratios 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 Surrogate threshold effect	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 = 0.99x0.99=0.98] log HR_{OS}=0.003+0.81xlog HR_{PFS} OS expected to be significant for HR_{PFS} of 0.86

Reference	Data type	Relationships analysed	Methods used	Results
Tang 2007[27]	Summary data from trials	Median PFS/TTP and OS,	Nonparametric Spearman rank correlation	Median PFS and OS: ρ=0.79 (95% CI, 0.65 to 0.87), p<0.00001 Median TTP and OS ρ=0.24 (95% CI,-0.13 to 0.55), p=0.21
		Differences (Δ) in median OS, PFS and TTP,	Nonparametric Spearman rank correlation coefficient	ΔPFS and ΔOS : ρ=0.74 (95% CI, 0.47 to 0.88), p<0.00001 Slope 1.02 (SE=0.16), R^2=0.65 ΔTTP and ΔOS : ρ=0.52 (95% CI, 0.004 to 0.81), p<0.05
		PFS and OS risk reduction	Linear regression (through origin) analysis	Risk reduction: Slope = 0.54 ± 0.10
Bowater 2008[10]	Summary data from trials	%gain in median TTP and % gain 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 b) P<0.005 for colorectal and p<0.001 for three other disease areas
Burzykowski 2008[15]	Individual patient level data	Individual level: PFS, TTP and OS	Spearman's rank correlation coefficient for correlation between endpoints	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)
		Trial level: HRs for PFS, TTP and OS	Spearman's rank correlation coefficient for treatment effects (HR) on endpoints	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)
			Linear regression for treatment effect (logHR)	Regression parameters not reported
Hotta 2009[26]	Summary data from trials	HR for TTP and OS (HR defined as ratio of medians)	Linear regression on HRs for TTP and OS Multivariate linear regression on HRs for TTP and OS (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

Reference	Data type	Relationships analysed	Methods used	Results
Miksad 2008[23]	Summary data from trials	HRs for PFS and OS	Kappa tests for agreement in direction of effects (HRs) Fixed effects linear regression for LogHRs (weighted by sample size)	Anthracyclines: Kappa=0.71 (0.36-1.00, p=0.0029) Taxanes: Kappa = 0.75 (0.42-1.00, p=0.0028) Anthracyclines: R^2=0.49, p=0.0019 Taxanes: R^2= 0.35, p=0.012
Sherrill 2008[11]	Summary data from trials	Treatment effects for TTP/PFS and OS (HR-1) Significance of treatment effect in TTP/PFS and OS	Linear model (through origin) on treatment effect, weighted by sample size Unweighted Pearson correlation between Hazard ratios Kappa test for agreement regarding significant treatment effect	Slope = 0.32 (95% CI: 0.20, 0.43), R^2 = 0.30 R=0.46 Kappa=0.47, P<0.05
Wilkerson 2009[21]	Summary data from trials	Differences in median PFS and OS HRs for PFS and OS	Linear regression Linear regression	Slope = 1.214, R^2=0.49, p<0.0001 R^2= 0.62, p<0.0001
Foster 2010[17]	Individual patient level data	Individual level: PFS status at 2,4,6mths and OS Trial level: LogHRs by trial centre (32 units) for PFS and OS	Individual: Multivariate landmark analysis for OS by PFS at 2,4,6 mths and c-index Trial level: Weighted least square regression, Spearman correlation coefficient, bivariate survival model (Copula)	Individual: 2mth: 0.40 (95%CI 0.30-0.52), c-index=0.60 4mth: 0.42 (95%CI 0.35-0.51), c-index=0.63 6mth: 0.41 (95%CI 0.35-0.49), c-index =0.65 Trial level: WLS R^2 = 0.79, Spearman ρ = 0.75 Copula R^2=0.80

Reference	Data type	Relationships analysed	Methods used	Results
Bowater 2011[28]	Summary data from trials	% gain in median TTP and % gain in PPS (PPS = median OS – median TTP)	Spearman's rank correlation for %change, Hypothesis (sign) test for proportion of trials with a) PPS%gain < TTP%gain, b) PPS%gain <0.5TTP%gain	P= 0.37 and 0.11 by tumour type (breast, colorectal) a) p<1% for both tumour types b) p<1% for both tumour types
Hotta 2011[25]	Summary data from trials	Median OS, median PFS and PPS (PPS= median OS – median PFS)	Linear regression analysis weighted by trial size	R^2=0.2563 for median OS and median PFS R^2=0.8917 for median OS and SPP
Halabi 2009[16]	Individual patient level data	Individual patient data on PFS and OS	Landmark analysis for OS by PFS at; 3mths 6 mths Kendalls tau for PFS and OS	 3mth PFS: HR = 2.0 (95% CI, 1.7 to 2.4; P<.001), 6mth PFS: 1.9 (95% CI, 1.6 to 2.4; P<.001), 0.30 (bootstrap SE = 0.0172, 95% CL = 0.26 to 0.32, p<0.00001).
Heng 2011[20]	Individual patient level data	Individual patient data on PFS and OS	Landmark analysis of OS by PFS at; 3mths 6 mths Kendalls tau for PFS and OS Correlation using Fleischer model	 3mth: 3.05 (95% CI, 2.42-3.84) 6mth: 2.96 (95% CI, 2.39-3.67), 0.42 (bootstrap SE, 0.016; 95% CI, 0.39-0.45; P < .0001) 0.66 (bootstrap SE, 0.025; 95% CI, 0.61-0.71)
Polley 2010[18]	Individual patient level data	Individual patient data on PFS and OS	Landmark analysis for OS by PFS at; 10weeks 18weeks 26 weeks	 10wk: 3.55 (95%CI, 2.28–5.52) 18wk: 2.06 (95%CI, 1.43–2.99) 26wk: 1.99 (95%CI, 1.38–2.85) (combined across all trials)

Reference	Data type	Relationships analysed	Methods used	Results
Mandrekar 2010[19]	Individual patient level data	Individual patient data on PFS and OS	Landmark analysis for OS by PFS at; 8weeks 12weeks 16weeks 20weeks 24 weeks	 HR=0.45 (95%CI 0.33-0.62) p<0.0001, c-index=0.63 HR=0.39 (95%CI 0.28-0.52) p<0.0001, c-index=0.67 HR=0.49 (95%CI 0.36-0.65) p<0.0001, c-index=0.66 HR=0.41 (95%CI 0.30-0.55) p<0.0001, c-index=0.68 HR=0.41 (95%CI 0.30-0.57) p<0.0001, c-index=0.68
Chirila 2011[13]	Summary data from trials	Median PFS/TTP and OS	Pearson product-moment correlation	PFS: 0.89 (95%CI 0.83 – 0.93) TTP: 0.75 (95%CI 0.59 – 0.84) PFS/TTP: 0.87 (95%CI 0.82-0.91)
			Spearman's rank correlation	PFS: 0.78 (95%CI 0.66 – 0.85) TTP: 0.59 (95%CI 0.37 – 0.74) PFS/TTP: 0.76 (95%CI 0.67-0.82)
		HR for PFS/TTP and OS (HR defined as ratio of medians)	Weighted least squares regression (weighted by trial size)	PFS/TTP: Slope = 0.41 (95%CI 0.30-0.52), R^2=0.48 PFS: Slope = 0.49 (95%CI 0.35-0.64), intercept = 0.52, R^2=0.59 TTP: Slope = 0.31 (95%CI 0.12-0.49), intercept =0.71, R^2=0.32 AUC = 0.795 (p<0.01)
			Diagnostic evaluation of regression equations (ROC curves for outcome of $HR_{os} \leq 0.8$)	$HR_{PFS} \leq 0.78$ has sensitivity =0.89 and specificity=0.69

OS=overall survival, PFS=progression free survival, TTP=time to progression, SPP=survival post-progression, RCT=randomised controlled trial, HR=hazard ratio, AUC=area under the curve, ROC=receiver operating characteristic, SE=standard error,

Table 2 summarises the methods used within the individual studies to analyse the relationship between PFS/TTP and OS. The analyses conducted fall into three broad categories. Four studies used aggregate data from multiple trials to examine the relationship between PFS and OS for individual trial arms. The methods for these studies are summarised in Section 4.1 with results presented in Section 4.3. Seven studies use IPD to examine the relationship between PFS and OS for individual patients. The methods for these studies are summarised in Section 4.2 and the results in Section 4.3. Thirteen studies examined the trial level relationship between treatment effect on PFS and treatment effect on OS with three using IPD and the remaining using aggregate data from multiple trials. These are summarised in Section 5. The majority of papers examining trial level association between outcomes report either rank correlation coefficients or linear regressions. Landmark analysis,[29] which assesses the prognostic impact on future survival of being alive and progression free at various time points, has been used by many of the more recent papers which use IPD to examine the association between outcomes at the individual level.

4. STUDIES EXAMINING THE RELATIONSHIP BETWEEN OUTCOMES FOR INDIVIDUALS OR SINGLE TREATMENT ARMS

4.1. METHODS IN STUDIES USING AGGREGATE DATA FROM MULTIPLE TRIALS

Four papers reported the relationship between median PFS or TTP and median OS using aggregate data from multiple trials.[13,22,25,27] Due to inconsistency in the definition of TTP, Chirilia *et al*[13] conducted separate analysis for PFS, TTP and a composite outcome including both PFS and TTP. One study also examined the relationship between OS and survival post-progression which was defined as the difference in median PFS and median OS.[25] Three papers report spearman ρ correlation coefficients[13,22,27] and two report linear regressions[22,25] with the more recent paper weighting the regression by trial size. Three papers included a variety of first-line treatments and one included first, second and third-line treatments.[13] One paper[25] conducted subgroup analysis to assess whether the relationship varied according to the following factors; treatment type (platinum-based, molecular-targeted agents, old chemotherapeutic agents), whether OS was the primary endpoint, the year of trial initiation and whether the publication was full text or abstract only. The other three papers did not conduct any analysis to assess whether factors such as treatment type had an effect on the relationship between PFS/TTP and OS.

4.2. METHODS IN STUDIES USING INDIVIDUAL PATIENT DATA

Four papers used IPD to estimate the correlation between PFS and OS.[14-16,20] In Buyse *et al*[14] the rank correlation coefficient ρ between PFS and OS was estimated "through a Hougaard bivariate copula distribution of these end points over the entire time range, or using the Kaplan-Meier estimates of PFS at 6 months and OS at 12 months." In Burzykowski *et al*[15] the association between PFS or TTP and survival was quantified through Spearman's rank correlation coefficient. Two papers assessed the "nonparametric Kendall tau rank correlation for bivariate censored data"[16,20] and one assessed the correlation between PFS and OS using the Fleischer model.[20]

Landmark analysis was used in the five papers to assess the prognostic impact of being alive and progression-free at various time points on future survival.[16-20] In the landmark analysis, multivariate Cox proportional hazards models were constructed for OS and these were stratified by progression-free status at various time points. Hazard ratios are reported for survival in patients who were alive and progression free at these time points compared to those who were not. Three of the models were stratified by trial protocol[16,17,19] while one reported separate analyses for each trial and a combined analysis adjusted for study protocol.[18]

All of the models were adjusted for multiple known prognostic factors, such as age, gender and performance status. Halabi *et al*[16] also conducted a subgroup analysis for trial participants who received docetaxel on the basis that this was the only known agent to prolong survival in castrate resistant prostate cancer. Halabi *et al* divided their sample into a training and testing set and applied the estimates of PFS to the testing data set and estimated misclassification error rates. Two papers evaluated model discrimination using the concordance index (c-index).[17,19] The c-index computes the probability that, for a pair of randomly chosen comparable patients, the patient with the higher risk prediction (e.g progression free at 3 months) will experience an event (e.g death) before the lower risk patient (e.g progressed before 3 months). A completely random prediction would have a c-index of 0.5, and a perfect rule will have a c-index of 1.0.[19]

4.3. RESULTS

4.3.1. Results from studies using aggregate data from multiple trials

In studies using aggregate data from multiple trials, the size of the correlations between the surrogate outcomes of PFS/TTP and the final outcome of OS were similarly wide ranging (0.24 to 0.89). Whilst the range of correlation coefficients reported for TTP (0.24 to 0.75) and PFS (0.481 to 0.89)

were overlapping, the values were consistently lower for TTP in those studies reporting both outcomes.[13,27]

Louvet *et al* reported a regression slope of 0.68, suggesting that each month of PFS is associated with 21 days of additional OS.[22] Hotta *et al* did not report the slope but their regression equation suggests that 26% of the variation in median OS is accounted for by the variation in PFS.[25] Subgroup analysis in one study found that treatment with platinum-based therapies was found to be a significant factor in the association between median OS and median PFS. None of the other factors considered in the subgroup analysis were found to have significant interaction terms.[25] Hotta *et al* also reported the association between survival post-progression and median OS which were strongly correlated (r^2=0.89), although this is to be expected given that they defined survival post-progression as the difference in median OS and median PFS.[25]

4.3.2 Results from studies examining individual level correlations

In studies using IPD data to examine individual level correlations, the size of the correlation coefficient for PFS and OS varied from 0.30[16] to 0.82,[14] although the variation is likely to be strongly related to the method used, as studies that reported more than one method also showed significant variation between estimates.[14,20] The one study reporting the correlation between TTP and PFS at the individual level gave an estimate of 0.682 which was similar to their estimate for PFS.

All of the landmark analyses conducted showed significant hazard ratios suggesting that people who are alive and progression free at the time points considered have a lower risk of subsequent mortality. Model discrimination as determined by the c-index was good (0.60 to 0.68) in all cases where it was estimated.

5. STUDIES EXAMINING THE RELATIONSHIP BETWEEN THE TREATMENT EFFECT FOR PFS/TTP AND THE TREATMENT EFECT FOR OS ACROSS MULTIPLE RCTS

Thirteen studies examined the relationship between the treatment effect on PFS or TTP and the treatment effect on OS.[9-11,13-15,17,21,23,24,26-28] Three studies used IPD to estimate measures of trial-level surrogacy[14,15,17] with the remaining papers using aggregate date from multiple trials. The analysis by Foster *et al*[17] used IPD data from three RCTs but as this was a small number they treated each trial centre as a separate unit giving 32 points for analysis.

5.1. DEFINITION OF TREATMENT EFFECT

Treatment effect was defined in several ways with three papers using the difference in median time-to-event (e.g median OS),[9,21,27] two using the proportional increase in median time-to-event[10,28] and 10 using hazard ratios (HR).[11,13-15,17,21,23,24,26,27] One paper defined the treatment effect as the hazard ratio minus unity (HR-1)[11] and another examined the "percent risk reduction" based on the HR.[27] Of the remaining papers using the HR, all but three[13,21,26] transformed the hazard ratio onto a log scale for the linear regression. Five papers defined the hazard ratio as the ratio of the median time-to-event between the trial arms,[11,13,23,24,26] which is consistent with assuming that the survival curve is exponential, although no justification was given for this assumption. Three papers estimated the HRs directly from IPD.[14,15,17] All three used a proportional hazards model to estimate the hazard ratios for PFS and OS. One specified that a joint model, based on the Hougaard copula, was used to estimate trial-specific treatment effects on PFS or TTP and survival by using marginal proportional hazards models with normally distributed, random trial–specific treatment effects for the TTP and survival.[15]

5.2. REGRESSION METHODS

Not all of the papers reported that the regression analyses were weighted according to trial size.[14,15,21,27] Two studies forced the intercept of the regression to zero[11,27] but both considered and discounted a non-zero intercept in exploratory analyses. One study explored the possibility of a non-linear regression by adding quadratic terms.[11] One study examined residual versus predicted plots and did diagnostic tests for normality and heteroscedasticity (non-constant error variance) to assess consistency with the assumptions of linear regression.[9]

One study evaluated the normality assumption, and presence of outliers or influence points using diagnostic tests and plots.[13] 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.[10,21,24,26] One study down-weighted the sample size by the number of trial arms to adjust for multiple comparisons.[23] One study included the comparison with the greatest treatment effect in the analysis and excluded all others[9] whilst others chose one comparison from each trial either at random[27] or by using clinical judgement (without regard for the size of correlation).[13]

Three studies[9,13,26] used multivariate analysis to explore whether any other factors were significant predictors of OS. Johnson *et al*[9] examined patients' age (trial median), performance status, stage of

disease (for lung cancer), year of trial (which was a surrogate of improvements in general medical care), trial quality, and use of rescue (or salvage) treatment. Hotta *et al*[26] explored the following six additional factors: year of trial initiation; use of cisplatin, carboplatin, and old agents; number of agents combined (combination therapy versus single agent therapy); number of randomized patients; and proportion of male patients. Chirila *et al*[13] examined the following factors using covariate analysis: line of therapy, performance status, clinical trial phase, crossover after progression, drug therapy, publication year, and median OS for the control group. Significant factors were then considered in subgroup analyses. They also conducted a subgroup analysis using just those studies that reported the HR in order to establish whether the ratio of medians is a good approximation for the HR. Miksad *et al*[23] refitted the analysis with interaction terms for two proxies which aimed to capture the impact of treatments given after the trial regimen. These were the year of last patient entry (before or after 1990) and line of trial therapy (first versus subsequent-line).

Several studies used subgroup analysis to assess whether the relationship between PFS/TTP and OS varied for trials with particular characteristics. The subgroups considered are summarised in Table 3.

Two studies analysed whether the relationship was different in trials with a reduced risk of bias by examining factors related to methodological quality. Sherrill *et al*[11] examined indicators for blinding and the availability of intention to treat analyses as markers for study quality. Johnson *et al*[9] assessed the quality of trials by use of the Schulz criteria and rated quality according to a 0–7 scale (low quality <3).

In one paper, the regression was validated by using it to predict OS treatment effects from PFS treatment effects in three validation trials.[14] Two papers used “leave-out-one” cross validation to predict the OS hazard ratio from the PFS hazard ratio for each trial using a regression fitted to all the remaining trials.[13,23]

The surrogate threshold effect, which is the minimum difference between surrogates required in order to predict a significant difference in OS, was reported in two papers. One paper reported the minimum difference required in the median PFS[9] and in the other reported the minimum hazard ratio required.[14] One paper used a ROC (Receiver operating characteristic) curve to whether various magnitudes of treatment effect for PFS are predictive of a clinically meaningful treatment effect in OS.[13]

5.3. OTHER MEASURES OF CORRELATION BETWEEN THE TREATMENT EFFECTS

In addition to linear regression methods, several papers reported other measures of correlation between treatment effects such as spearman rank correlation coefficients.[15,17,27] Foster *et al*[17] also reported a formal trial-level surrogacy measure, known as the Copula R.

Two papers examined the relationship between the percentage gain or loss in median post-progression survival (PPS) and the percentage gain or loss in TTP.[10,28] PPS was defined as the difference between median OS and median TTP. They used a sign test to examine whether the percentage gain (or loss) in median post-progression survival was greater or less than either the percentage gain in median TTP or half the percentage gain in median TTP. In both these papers, trial comparisons reporting exactly no difference in median TTP were excluded from the analysis. In Bowater *et al*,[28] a sensitivity analysis was also conducted excluding those studies where the difference in TTP was non-significant. A spearman correlation test was also used to look for correlations in the percentage change in PPS and TTP.

Table 3: Subgroups considered

Paper	Factors analysed
Hackshaw *et al*[24]	• 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*[11]	• 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
Miksad *et al*[23]	• Strict PFS definition, • Year last patient recruited • First / subsequent line treatment
Hotta *et al*[26]	• 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

5.4. RESULTS

5.4.1 Regression parameters

In the studies using aggregate data from multiple trials, the R^2 from linear regression varied from 0.19 to 0.56 for the treatment effect in TTP and OS, and 0.35 to 0.65 for the treatment effect in PFS and OS. For studies defining the treatment effect in terms of absolute change in survival between treatments arms, the R^2 varied from 0.19 to 0.65. For studies defining the treatment effect using the hazard ratio, the R^2 varied between 0.30 and 0.59. When considering the R^2 for treatment effect by tumour type, the values were fairly consistent for three studies in breast cancer (0.30 to 0.56), and similarly for three studies in colorectal cancer (0.32 to 0.65).

The two highest R^2 values were reported by the studies using IPD. Foster *et al*[17] reported an R^2 of 0.79 between PFS and OS when using log HRs as the treatment outcome in extensive small-cell lung cancer. Buyse *et al*[14] reported an R value much higher than the values reported in other studies ($R = 0.99$), but they gave a 95% CI spanning unity, which suggests that this estimate should be treated with caution.

The slope parameters in the linear regression analyses ranged from 0.32 to 0.81 for studies using the hazard ratio as the measure of treatment effect. This suggests that reductions in the hazard for PFS/TTP may translate into smaller reductions in the hazard for OS. For studies defining the treatment effect using the difference in median time-to-event, the slope parameter varied more significantly and ranged from 0.0961 to 1.214. The slope parameters appear to be higher in the studies using PFS rather than TTP to predict OS.

5.4.2 Modifying factors examined in multivariate or subgroup analysis

None of the factors considered in the multivariate analyses by Johnson *et al*[9] or Hotta *et al*[26] were found to be significantly predictive of survival. Chirila *et al*[13] found that line of therapy had a significant interaction effect (p=0.03) and subgroup analysis found a higher R^2 (0.54 vs 0.37) in the studies examining first-line therapy compared to those examining second-line therapy. In Miksad *et al,*[23] there was a statistically non-significant interaction between line of trial treatment and the hazard ratio for PFS when predicting the hazard ratio for OS. In the subgroup analyses, Sherrill *et al*[11] found a particularly strong correlation (R^2=0.93) in the four studies with only HER2+ patients and a higher slope, albeit with very wide confidence intervals, in the 12 studies of hormonal therapies. None of the other factors considered in subgroup analysis by Sherrill *et al* had a significant impact on the regression. Hotta *et al*[26] report that none of the study characteristics assessed in subgroup analysis seemed to affect the association between the hazard ratios for TTP and OS. Hacksaw *et al*[24] found that the regression slopes were not significantly different (p=0.15) for the studies that recruited patients before 1990 when second line therapies were not commonly used and the results were similar in the 9 additional studies that reported PFS rather than TTP. Chirilia *et al*[13] found that the R^2 was higher when using the logHRs rather than the ratios of the means, however, the results were similar for the log of the ratio of medians when the analysis was restricted to the same subgroup of studies.

5.4.3 Validation of predictive relationships

The three studies which attempted to validate the predictive strength of their regression had varying results. Buyse *et al*[14] found that the actual hazard ratios were within the prediction limits in all three studies and the prediction limits fell within the CI of the actual hazard ratios in two studies. In Miksad *et al,*[23] all but two of the 34 observed estimates fell within the prediction limits, but these were wide and crossed the line of equivalence in all cases, with the predicted midpoint sometimes falling on the opposite side of the line from actual midpoint. Chirila *et al*[13] found that the number of over and under predictions from the cross validation was approximately equal. Their ROC analysis

found the HR for PFS was significantly better than chance ($p<0.03$) at predicting a HR for OS of 0.80 or better.

5.4.4 Correlation coefficients

Two studies reported a spearman rank correlation coefficient varying from 0.48[15] to 0.75[17] for treatment effects in PFS and OS based on the log hazard ratio whilst a third reported a coefficient of 0.74[27] based on the difference in median PFS and OS. Foster *et al*[17] reported fairly consistent measures of correlation varying from 0.75 to 0.80 using three separate methods in the same data set.

5.4.5 Indirect information based on the relationship between TTP and PPS

The two papers[10,28] which examined the relationship between TTP and post-progression survival found that changes in PPS are small relative to changes in TTP. Whilst this doesn't provide direct information on the relationship between TTP and OS, it does suggest that changes in the absolute TTP will translate into gains in OS. Although Broglio *et al*[30] have showed using simulation that significant differences in PFS may not translate into significant differences in OS even when there is no difference in PPS between arms, purely due to variability in PPS diluting the OS comparison.

Table 4: Summary of results by tumour type and level of evidence (Elston and Taylor 2009[5])

Tumour type	Number of papers	Level 2 evidence [Association between surrogate outcome and OS]	Level 1 evidence [Association between treatment effects for surrogate and OS]
Colorectal[9,13,14,22,27]	5	PFS: ρ for individual outcomes in 1 paper[14] (0.82, p <0.05) and study level median in 3 papers[13,22,27] (0.481 to 0.79, p<0.05 for all three estimates) Slope for median outcomes = 0.68 from one paper[22] TTP: ρ for study level median outcome from two papers[13,27] (0.24 , p>0.05 and 0.59, p<0.05)	PFS: ρ =0.74 (p<0.05), R^2 of 0.65, slope = 1.02 for difference in median from one paper[27] R^2 =0.98 and slope =0.81 for logHR (p values not reported) from one paper[14] slope = 0.54 for risk reduction in one paper[27]. R^2 = 0.59, slope = 0.49 for HR (ratio of medians) from one paper[13] TTP: R^2 of 0.33 (p<0.05), Slope =0.0961 for difference in median from one paper[9] ρ=0.52 (p<0.05) for difference in median from one paper[27] R^2 = 0.32, slope = 0.31 for HR (ratio of medians) from one paper[13]
Breast[11,15,23,24]	4	PFS: ρ=0.688; (p<0.05) for individual outcomes in 1 paper[15] TTP: ρ=0.682; (p<0.05) for individual outcomes in 1 paper[15]	PFS: ρ= 0.48 (p>0.05) for log HR from one paper[15] Anthracyclines: R^2=0.49, p<0.05 for log HR from one paper[23] Taxanes: R^2= 0.35, p<0.05 for log HR from one paper[23] TTP: R^2=56%, (P<0.001), slope = 0.5082 for logHR from one paper[24] ρ= 0.49 (p>0.05) for log HR from one paper[15] PFS/TTP (combined): Slope = 0.32 (p<0.05), R^2 = 0.30 for(HR-1) from one paper[11]

Non-small cell lung	4[9,19,25,26]	PFS: R^2=0.2563 for medians from one paper[25] Significant hazard ratios for 8,12,16, 20 and 24 weeks in landmark analysis.[19] TTP: None	PFS: None TTP: R^2=0·19; p<0.05, slope=0·616 for gain in median from one paper[9] R^2 = 0.33, p<0.01 for HR from one paper[26]
Extensive small cell lung	1[17]	PFS: Significant hazard ratios for 2, 4 and 6mths in landmark analysis.	PFS: Correlations coefficients for LogHR ranging from 0.75 to 0.80 depending on measure used (p values not reported)
Prostate	1[16]	PFS: Significant hazard ratios for 3mths and 6mths in landmark analysis. Kendall's tau = 0.30 (p<0.05)	
Brain (Glioblastoma multiforme)	1[18]	PFS: Significant hazard ratios for 10, 18 and 26 weeks in landmark analysis.	
Renal cell	1[20]	PFS: Significant hazard ratios for 3mths and 6mths in landmark analysis. Significant correlations using Kendall's tau and Fleischer's model (0.42 and 0.66 respectively)	

6. DISCUSSION

6.1. CONSISTENCY OF METHODS USED

This review demonstrates that a wide variety of methods have been used to examine the relationship between PFS/TTP and OS. Broadly these fall into several categories. Some papers have used IPD to establish surrogacy at either the individual level or the trial level, some have examined the association between treatment effects using aggregate data, and others have examined the association between outcomes for individual trial arms at the aggregate level. Even within these broad categories a variety of techniques have been used. The lack of a standardised approach makes it difficult to establish whether there is a consistent relationship between PFS/TTP and OS. However, it is likely that the relationship will not be constant across different types of tumour, and may also vary according to factors such as class of treatment being used and the effectiveness of subsequent-line therapies.

6.2. LEVEL OF EVIDENCE IN DIFFERENT TUMOUR TYPES

The evidence for PFS and TTP as surrogate outcome for OS in different tumour types is summarised in Table 4 according to Elston and Taylor's levels of evidence.[5] In some cancer types, such as colorectal cancer and breast cancer, evidence has been identified examining both trial level and individual level surrogacy, with trial level surrogacy being assessed using both IPD and aggregate trial data. In other cancers types such as metastatic renal cell carcinoma and primary malignant brain tumour, only individual level surrogacy has been assessed.

In colorectal cancer, there is consistent level 1 and 2 evidence for PFS as a surrogate for OS, but not for TTP as a surrogate of OS, where the correlation coefficient was non-significant for median TTP, but significant for the treatment effect on median TTP. In breast cancer, the rank correlation coefficient was significant for both PFS and TTP at the individual level but not for the treatment effect on PFS and TTP at the trial level. In non-small cell lung cancer, there was level 1 evidence for PFS and level 2 evidence for TTP showing a significant correlation in both cases.

6.3. LIMITATIONS OF THE EVIDENCE AVAILABLE

When using aggregate data from multiple studies, it can be difficult to determine all of the required information from study reports. For example, Sherrill *et al*[11] found that the definitions of PFS and TTP were so variable between studies, that they decided to combine the outcomes. Furthermore,

many of the papers included in this review estimated the hazard ratios from the ratio of median survival in the treatment and control arm, because hazard ratios were not consistently reported in the trials. This definition of the hazard ratio implicitly assumes that the survival curves are exponential and therefore that the hazard is constant. These problems can be avoided if IPD are available from multiple trials as a consistent analysis can be framed using all the available data. Such an analysis would also allow the within patient correlation between outcomes to be taken into account. However, the availability of IPD may limit the number of trials that can be included in such analyses by any one investigator and therefore limit their external validity.

The external validity of any relationship is likely to be higher if a greater number of trials are included in the analysis. However, including a broader range of trials will inevitably lead to increased heterogeneity in the trial populations. Very few papers attempted to examine the impact of confounding factors such as differences in the patient population between trials or the risk of bias, which is known to affect treatment effect in RCTs. However, in those papers that did examine confounding factors, few were found to be significant.

Many of the papers included trials with a wide variety of treatments in the analysis. Only a few papers attempted to determine whether the relationship varied by treatment class and the results were not consistent across these studies.[11,13,23,25,26] Some tried to assess whether the availability of subsequent therapies may have affected the relationship by considering trials conducted before second-line therapies became available separately from more recent trials.[23,24]

6.4. LIMITATIONS OF THIS REVIEW

One of the limitations of this review is that we were unable to conduct a systematic search and therefore we cannot claim to have identified all of the relevant papers in any particular tumour type. A more systematic search could have been conducted if we had narrowed our scope to a single tumour type. However, by keeping our scope broad we have been able to examine the variety of methods used.

Another limitation is that we have not attempted to replicate any of the analyses presented in the included papers. This may be useful, as it would allow us to examine whether all of the assumptions made in the presented analyses are supported by the primary data and whether the conclusions change when all relevant trials are considered in a single analysis and after updating for any recently

published trials. However, any re-analysis that is based on aggregate trials outcomes rather than IPD will continue to be hindered by the limitations described earlier.

The majority of the literature concerning surrogate outcomes is interested in identifying surrogate outcomes that either reduce the length of follow-up required in an RCT, or which avoid potential bias in measures such as OS which may be introduced by subsequent therapies. None of the statistical analyses reviewed here have been conducted with the specific aim of determining the exact relationship between PFS and OS that can be reliably assumed within a cost-effectiveness model. Whilst some of the papers attempted to use their linear regressions to predict treatment effect in OS from treatment effect in PFS for a set of validation studies, these sometimes lead to misleading predictions regarding the size and statistical significance of the treatment effect for OS.[14]

7. CONCLUSIONS

We have found that the level of evidence available supporting a relationship between PFS/TTP and OS varies considerably by cancer type and is not always consistent even within one specific cancer type. Furthermore, even where robust consistent evidence supporting a correlation between the treatment effects (i.e level 1 evidence according to Elston and Taylor) is available, it is unclear how that should be converted into a quantified relationship between PFS and OS treatment effects within a cost-effectiveness model. Therefore, any cost-effectiveness analysis which makes a strong assumption regarding the relationship between PFS and OS should be treated with caution. We would support Elston and Taylor in recommending that any cost-effectiveness analysis based on a surrogate relationship between PFS and OS should be supported with a transparent explanation of how the relationship is quantified in the model and should be accompanied by sensitivity analysis exploring the uncertainty associated with that relationship and a systematic review of papers examining the relationship between PFS and OS in the relevant setting. This would allow decision makers to judge the appropriateness of the model in light of the evidence available in that specific disease area.

8. REFERENCES

1. Therasse, P., Arbuck, S.G., Eisenhauer, E.A., Wanders, J., Kaplan, R.S., Rubinstein, L. et al. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *Journal of the National Cancer Institute* 2000; 92(3):205-216.

2. Tappenden, P., Jones, R., Paisley, S., Carroll, C. Systematic review and economic evaluation of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer. *Health Technology Assessment* 2007; 11(No.12).

3. Roche Products Limited. Rituximab for the 1st line treatment of chronic lymphocytic leukaemia, Roche submission to the National Institute for Health and Clinicial Excellence. 2008; available from http://www.nice.org.uk/nicemedia/live/12039/43581/43581.pdf (accessed Oct. 2010).

4. Dundar, Y., Bagust, A., Hounsome, J., McLeod, C., Boland, A., Davis, H., Walley, T., Dickson, R. Rituximab for the first-line treatment of stage III/IV follicular non-Hodgkin's lymphoma. 2006; available from http://www.nice.org.uk/nicemedia/live/11591/46698/46698.pdf (accessed Nov. 2010).

5. Elston, J., Taylor, R.S. Use of surrogate outcomes in cost-effectiveness models: A review of United Kingdom health technology assessment reports. *International Journal of Technology Assessment in Health Care* 2009; 25(1):6-13.

6. Bucher, H.C., Guyatt, G.H., Cook, D.J., Holbrook, A., McAlister, F.A. User's guides to the medical literature: How to use an article measuring the effect of an intervention on surrogate end points. *JAMA: The Journal of the American Medical Association* 1999; 282(8):771-778.

7. Lassere, M.N., Johnson, K.R., Boers, M., Tugwell, P., Brooks, P., Simon, L. et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *The Journal of Rheumatology* 2007; 34(3):607-615.

8. Fleming, T.R. Surrogate endpoints and FDA's accelerated approval process. *Health Affairs* 2005; 24(1):67-78.

9. Johnson, K.R., Ringland, C., Stokes, B.J., Anthony, D.M., Freemantle, N., Irs, A. 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. *Lancet Oncology* 2006; 7(9):741-746.

10. Bowater, R.J., Bridge, L.J., Lilford, R.J. The relationship between progression-free and post-progression survival in treating four types of metastatic cancer. *Cancer Letters* 2008; 262(1):48-53.

11. Sherrill, B., Amonkar, M., Wu, Y., Hirst, C., Stein, S., Walker, M. 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(10):1572-1578.

12. Saad, E.D., Katz, A., Buyse, M. Overall Survival and Post-Progression Survival in Advanced Breast Cancer: A Review of Recent Randomized Clinical Trials. *Journal of Clinical Oncology* 2010; 28(11):1958-1962.

13. Chirila, C., Odom, D., Devercelli, G., Khan, S., Sherif, B., Kaye, J. et al. Meta-analysis of the association between progression-free survival and overall survival in metastatic colorectal cancer. *International Journal of Colorectal Disease* 2011;1-12.

14. Buyse, M., Burzykowski, T., Carroll, K., Michiels, S., Sargent, D.J., Miller, L.L. et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology* 2007; 25(33):5218-5224.

15. Burzykowski, T., Buyse, M., Piccart-Gebhart, M.J., Sledge, G., Carmichael, J., Luck, H.J. 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* 2008; 26(12):1987-1992.

16. Halabi, S., Vogelzang, N.J., Ou, S.S., Owzar, K., Archer, L., Small, E.J. Progression-Free Survival as a Predictor of Overall Survival in Men With Castrate-Resistant Prostate Cancer. *Journal of Clinical Oncology* 2009; 27(17):2766-2771.

17. Foster, N.R., Qi, Y., Shi, Q., Krook, J.E., Kugler, J.W., Jett, J.R. 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(6):1262-1271.

18. Polley, M.Y., Lamborn, K.R., Chang, S.M., Butowski, N., Clarke, J.L., 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(3):274-282.

19. Mandrekar, S.J.P., Qi, Y., Hillman, S.L., Allen Ziegler, K.L., Reuter, N.F., Rowland, K.M. et al. Endpoints in Phase II Trials for Advanced Non-small Cell Lung Cancer. *Journal of Thoracic Oncology* 2010; 5(1):3-9.

20. Heng, D.Y.C.M., Xie, W., Bjarnason, G.A., Vaishampayan, U., Tan, M., Knox, J. et al. Progression-free survival as a predictor of overall survival in metastatic renal cell carcinoma treated with contemporary targeted therapy. *Cancer* 2011; 117(12):2637-2642.

21. 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(5):379-385.

22. 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(11):2033-2038.

23. Miksad, R.A., Zietemann, V., Gothe, R., Schwarzer, R., Conrads-Frank, A., Schnell-Inderst, P. et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. *International Journal of Technology Assessment in Health Care* 2008; 24(4):371-383.

24. 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(11):1215-1221.

25. Hotta, K., Kiura, K., Fujiwara, Y., Takigawa, N., Hisamoto, A., Ichihara, E. 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(11).

26. Hotta, K., Fujiwara, Y., Matsuo, K., Kiura, K., Takigawa, N., Tabat, M. 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* 2009; 4(3):311-317.

27. Tang, P.A., Bentzen, S.M., Chen, E.X., Siu, L.L. 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* 2007; 25(29):4562-4568.

28. Bowater, R.J., Lilford, P.E., Lilford, R.J. 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(3):207-214.

29. Anderson, J.R., Cain, K.C., Gelber, R.D. Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. *Journal of Clinical Oncology* 2008; 26(24):3913-3915.

30. Broglio, K.R., Berry, D.A. Detecting an overall survival benefit that is derived from progression-free survival. *Journal of the National Cancer Institute (Cary)* 2009; 101(23):1642-1649.

9. APPENDICES

APPENDIX 1: EXCLUDED STUDIES

Reference	Reason
Saad E, Katz A, Hoff P, Buyse M. Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. Annals of Oncology 2010;21(1):7-12.	Systematic review reporting PFS and OS but not examining relationship between PFS and OS
Lee L. Identification of potential surrogate end points in randomized clinical trials of aggressive and indolent non-Hodgkin's lymphoma: correlation of complete response, time-to-event and overall survival end points.[Article]. Annals of Oncology 2011;22(6):1392-1403.	Treatment has curative rather than palliative intent
Broglio KR, Berry DA. Detecting an Overall Survival Benefit that Is Derived From Progression-Free Survival. Journal of the National Cancer Institute (Cary) 2009;101(23):1642-1649.	Simulated data

APPENDIX 2: SEARCH STRATEGIES

The initial citation search of the three relevant papers retrieved six further relevant papers providing nine relevant papers to inform the initial search. Exploratory searches on Medline were conducted and compared against the nine relevant papers. Line 33 shows the results of the initial strategy. This resulted in over 3000 records being retrieved , but as shown in line 34, all of the nine target papers were identified. Lines 36, 39, 42, and 45 show attempts to make the search more specific by omitting some of the terms, with the number of target papers identified given on the line below (lines 37, 40, 43 and 46).

Database: Ovid MEDLINE(R) <1946 to February Week 2 2012>

Search Strategy:

Searches 1 to 9 specify the references identified in the first citation search.

1 "9".fc_issue. and "7".fc_vol. and "johnson$".fc_auts. and "response rate or time to progression".fc_titl. and "2006".fc_pubyr. (1)

2 "bowater$".fc_auts. and "the relationship between progression".fc_titl. and "cancer lett$$".fc_jour. and "2008".fc_pubyr. (1)

3 "sherrill b*$".fc_auts. and "relationship between effects".fc_titl. and "british journal of cancer$".fc_jour. and "2008".fc_pubyr. (1)

4 "5".fc_issue. and "15".fc_vol. and "wilkerson$".fc_auts. and "Progression-Free Survival Is Simply ".fc_titl. and "Cancer Journal$".fc_jour. and "2009".fc_pubyr. (1)

5 "23".fc_issue. and "101".fc_vol. and "broglio$".fc_auts. and "detecting an Overall Survival Benefit ".fc_titl. and "Journal of the National Cancer Institute$".fc_jour. and "2009".fc_pubyr. (1)

6 "hotta$".fc_auts. and "Time to Progression ".fc_titl. and "2009".fc_pubyr. (1)

7 "buyse$".fc_auts. and "Progression-free survival ".fc_titl. and "2007".fc_pubyr. (1)

8 "saad$".fc_auts. and "Overall Survival and Post-Progression Survival".fc_titl. and "2010".fc_pubyr. (1)

9 "bowater$".fc_auts. and "estimating changes".fc_titl. and "2011".fc_pubyr. (1)

10 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 (9)

Search 11-12 are for cancer terms

11 *Neoplasms/ (188873)

12 (cancer$ or neoplasm$ or tumour$ or tumor$ or malignan$).ti. (904183)

13 11 or 12 (965136)

Search terms 14-21 are around progression free survival, disease progression post progression survival

14 Disease-Free Survival/ (32602)

15 progression free survival.tw. (10226)

16 pfs.tw. (4397)

17 time to progression.tw. (5537)

18 post progression.tw. (10)

19 postprogression.tw. (17)

20 spp.tw. (37581)

21 Disease Progression/ (81765)

22 or/14-21 (159532)

Searches 23-24 are for overall survival

23 overall survival.tw. (47963)

24 OS.tw. (21321)

25 or/23-24 (61215)

Searches 26-31 are for terms to indicate a relationship between the progression free survival and overall survival.

26 Regression analysis/ (94426)

27 regression.tw. (286119)

28 relationship.tw. (545559)

29 correlation.tw. (460727)

30 end point.tw. (19669)

31 survival analysis/ (86277)

32 or/26-31 (1293865)

Below are different combinations of the terms to indicate a relationship with numbers of results and after each 1 how many of the 9 relevant citations are picked up

33 13 and 22 and 25 and 32 (3477)

34 10 and 33 (9)

35 26 or 27 or 28 or 29 or 31 (1278150)

36 13 and 22 and 25 and 35 (3175)

37 10 and 36 (8)

38 26 or 27 or 28 or 29 (1206968)

39 13 and 22 and 25 and 38 (1395)

40 10 and 39 (7)

41 26 or 27 (337227)

42 13 and 22 and 25 and 41 (633)

43 10 and 42 (4)

44 26 or 27 or 28 (833269)

45 13 and 22 and 25 and 44 (935)

46 10 and 45 (5)