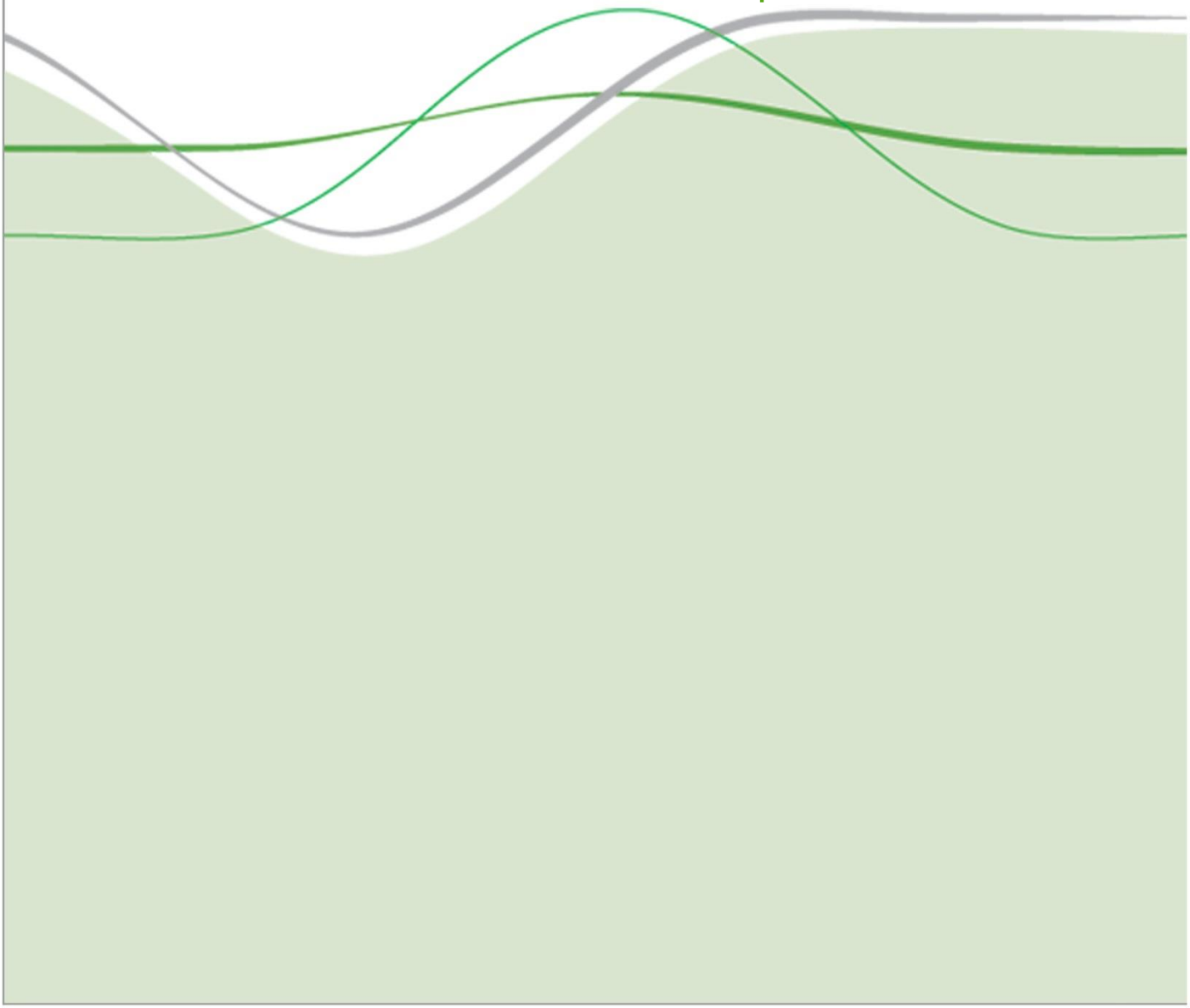
Office of Health Economics Research

Research Paper 16/07

Extrapolation from Progression-Free Survival to Overall Survival in Oncology

December 2016

Alastair Fischer, Karla Hernandez-Villafuerte, Nicholas Latimer and Christopher Henshall



Extrapolation from Progression-Free Survival to Overall Survival in Oncology

Alastair Fischer[a], Karla Hernandez-Villafuerte[a], Nicholas Latimer[b] and Christopher Henshall[c]

[a]Office of Health Economics [b]ScHARR, University of Sheffield [c]Independent consultant

December 2016

For further information please contact:

Alastair Fischer afischer@ohe.org

The Office of Health Economics (a company limited by guarantee of registered number 09848965)

Southside, 105 Victoria Street

London SW1E 6QT

United Kingdom

Tel: +44 207 747 8858

ABOUT OHE RESEARCH PAPERS

OHE Research Papers are intended to provide information on and encourage discussion about a topic in advance of formal publication.

Any views expressed are those of the authors and do not necessarily reflect the views or approval of OHE, its Editorial or Policy Committees, or its sponsors.

FUNDING AND ACKNOWLEDGEMENTS

The Pharmaceutical Oncology Initiative (POI) of the Association of the British Pharmaceutical Industry (ABPI) commissioned the Office of Health Economics (OHE) in April 2016 to undertake a landscape study on methods and approaches to extrapolation from clinical endpoints measured in trials to overall survival (OS). The POI group wished to understand the statistical and health economics issues around extrapolating from the kinds of endpoints measured in trials, such as progression-free survival (PFS), to overall survival. It wished to identify whether it could usefully invest in resolving issues in such a way that would, for example, change the way the National Institute for Health and Care Excellence (NICE) and other health technology assessment (HTA) bodies deal with an absence of OS data.

The terms of reference of the project were to:

a) Identify and explain what the statistical and health economics issues are with current approaches to extrapolating from clinical endpoints measured in trials to overall survival. The focus of the work will be on the extrapolation from progression-free survival to overall survival, as this is required for HTA submissions and calculation of quality-adjusted life years (QALYs) for cost effectiveness analysis.

b) Identify the principal weaknesses and gaps in current methods and approaches, where further research is required.

We gratefully acknowledge the contributions of Francesco Pignatti (EMA), Eli Gavraj and Ian Watson (NICE), Prof Andrew Stevens (Chair of a NICE Appraisal Committee), Oriana Ciani and Prof Rod Taylor (University of Exeter Medical School), and contributors to the workshop and draft report from POI member organisations.

Nevertheless, views expressed in this paper are those of the authors and are not necessarily those of the POI, European Medicines Agency (EMA), NICE or the University of Exeter.

AUTHORS' CONTRIBUTIONS

The literature review was led and written by KHV with input from AF. The Workshop was chaired by Nancy Devlin and led by AF. The report of the Workshop and the review of the two technical papers were drafted by AF, with input from KHV. NL gave advice on technical aspects and on how surrogacy has been treated by HTA bodies such as NICE. CH had significant editorial contributions on structure and themes of the report. The report was quality assured internally at OHE by Paula Lorgelly and by Prof Martin Buxton of the OHE Editorial Committee.

ABSTRACT

***Background**:* The outcomes from clinical and other healthcare trials of most interest to patients and health systems are usually increases in the quality and length of life (overall survival (OS)) as a result of treatment or other intervention. This poses a problem, because complete knowledge on the true increase in length of life is not available until the last person in the trial dies. However, if the increased length of survival is sufficiently highly correlated with an entity that is observable within the trial period or soon after the treatment has finished, and is also highly-enough correlated with the treatment, the observable entity can replace OS, or can be used to estimate OS, without much error. Such an entity is called a surrogate endpoint. The most widely-used surrogate endpoint for OS in oncology is progression-free survival (PFS). Some aspects of cancer treatment make the evaluation of the quality of surrogate endpoints such as PFS problematic (e.g. crossover among trial arms and multiple subsequent therapies). This document aims at: (1) analysing the methods used to extrapolate from PFS to OS in the field of oncology; (2) identifying whether a clear guidance exists in the literature about what is considered to be 'best practice' in extrapolation from PFS to OS; (3) evaluating whether the relationship between PFS and OS varies by tumour type; and (4) determining the key limitations, weaknesses and gaps in the current literature and methods used to test PFS surrogacy, where future research efforts might usefully be targeted.

***Methodology**:* We extend the literature review carried out by Davis, Tappenden and Cantrell (2012) from 2012 to 2016, using similar inclusion and exclusion criteria, we interview experts from regulatory and reimbursement bodies, we explore academic research into the methodology of surrogacy and the need for better reporting of surrogacy papers, and we report on a workshop in which we bring together representatives from these groups along with experts from POI member bodies. The literature review is semi-self-contained within the report.

***Results**:* A number of factors affect the relationship between PFS and OS. Therefore, there is no unique correct answer for the question of whether PFS is an appropriate surrogate for OS in oncology. Many of these factors are related to the length and characteristics of post-progression survival (PPS). In addition, the results suggest that co-integration and linear regression are the main methodologies applied to extrapolate from PFS to OS. Nevertheless, there exists high variation in the characteristics of the models used. These appear to hinder the correct comparison of results. Moreover, we have identified a lack of rigour in the application of linear regression methodology. Additionally, results indicate that very little research has been undertaken using individual patient data (IPD) from trials. Methodological advances (Buyse et al. (2016) have produced the concept of the surrogate threshold effect (STE), a correlation between the surrogate and OS above which a surrogate endpoint can ensure with sufficient confidence that a treatment will not reduce overall survival. A second advance by Stevens et al. (2014) outlines an economic approach in which the benefits of earlier adoption of a treatment through the use of a surrogate endpoint are balanced against the possibility that a decision to adopt a treatment using the surrogate endpoint may not be cost effective.

***Conclusion**:* Any consideration of evidence relating to PFS should consider both tumour type and other factors, particularly those related to post-progression survival (PPS). Protocols of future follow-up of clinical trial patients should specify procedures for

gathering information about the effect of post-progression management of the disease. This should allow stronger conclusions to be extracted from statistical analyses. Improved reporting standards will aid in achieving this goal. In addition, it is very likely that increasing the use of individual patient data (IPD) will result in greater precision in estimating the benefits of worthwhile drugs. Using only a single estimate for PFS and a single estimate for OS from a reported trial wastes an enormous amount of information and may result in later adoption of effective and cost effective drugs. More widespread knowledge of methodological advances in statistical and economic analysis may make significant improvements in the use of surrogate endpoints.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The outcomes from clinical and other healthcare trials of most interest to patients and health systems are usually increases in the length and quality of life as a result of treatment or other intervention. In many if not a majority of cases, the increase in length of life, that is overall survival (OS), is the most important factor in decisions on uptake.

This poses a problem, because complete information on the increase in length of life associated with a new treatment is not available until the last person in the trial dies. However, in reality, trial data are censored, with some patients not experiencing the event of interest (death). In practice, estimates of OS are required for economic evaluation and the most common approach for arriving at these estimates is to use survival analysis methods to extrapolate beyond the observed trial period. Parametric models, which make various assumptions about the underlying hazard and survival functions, are fitted to trial data and are used to extrapolate into the future. (Latimer 2013) However, post-progression survival data can be confounded by treatment switching, or receipt of various lines of post-study treatment. Thus in some circumstances OS data from the trial may not be appropriate for extrapolation purposes. An alternative to extrapolating OS data is to use progression-free survival (PFS) as a "surrogate" outcome. If the increased length of survival is sufficiently highly correlated with an entity such as PFS that is observable within the trial period or soon after the treatment has finished, and is also highly-enough correlated with the treatment, the observable entity can replace OS without much error. Such an entity is called a proxy, surrogate outcome, or surrogate endpoint.

The question is whether the risks of basing decisions on uptake of a drug on estimates of OS and quality of life through to death, derived from surrogate outcomes, outweigh the risks of basing decisions on direct measurements of OS and quality of life through to death. The risks of using surrogates are that we will over- or under-estimate the true OS and/or quality of life (QoL) benefits. If we over-estimate them and the drug is actually no better, or even worse, than current treatment options, health systems and patients adopting the drug will suffer a decrease in value from what they expected when they decided to use it (and in the case of it being worse than current treatments, an absolute decrease in value). If we underestimate OS and QoL benefits, patients and health systems not adopting the drug would lose whatever value it offers. The risks of using direct measurements are that the consequent delays lead to a loss of revenue for industry, and patients and health systems lose whatever benefit that the drug offers over existing treatments during the period of "delay".

It is therefore important to determine the circumstances within which the use of surrogate endpoints in decision-making can result in an overall improvement in patient health and in value for health systems, industry and society in general.

The most widely-used surrogate endpoint for OS in oncology is progression-free survival (PFS), which has been in use since the 1970s. Two aspects of cancer treatment make the evaluation of surrogate endpoints such as PFS particularly problematic. The first is that if in a trial it is sufficiently strongly believed that the treatment arm is outperforming the control arm, patients will often switch from the control arm to the treatment arm. This has the effect of reducing the size of the treatment effect estimated in a conventional intention to treat analysis. Statistical methods have been developed to adjust for the effects of treatment switching and these have been used in economic

evaluation. (Latimer et al 2014) However, switching adjustment methods are not perfect and treatment switching makes it harder to evaluate whether a surrogate endpoint improves OS, particularly when the surrogate is PFS and disease progression is used as the trigger for switching.

The second aspect is that cancer treatment is in most cases a succession of different treatments. Early cancer detection (before a cancer metastasises) can sometimes allow tumour removal by operation and a cure performed. But where this cannot be performed or when the cancer has spread, a first-line treatment will be employed and after a time will cease to work, and be replaced by a succession of subsequent treatments. In that case, there could be a PFS for several treatments but only one OS. The treatments thus confound each other and the relationship between one particular PFS and OS will not be known, except where a number of patients have gone through the same lines of treatment except for the last line. In that case, the relationship between PFS and OS for last line, using the penultimate line as the base, can be estimated without confounding.

A further challenge is that the treatment in any line is often a combination of drugs, and trials of new drugs are often therefore trials of adding that drug into the combination, or replacing one drug in the combination with it.

It may be argued that observed OS is a suitable measure of the true impact on OS of introducing a new drug into the treatment pathway, as in reality patients are likely to receive a sequence consisting of several drugs. Pragmatically, decision-makers are often interested in the effectiveness and cost-effectiveness of inserting a new treatment into a treatment pathway, and if a multitude of subsequent treatments means that the OS benefit is relatively small then that is what should be reflected in the economic evaluation. Direct treatment switching, where patients in the control group move onto the experimental treatment, is more problematic because this does not reflect a realistic treatment pathway. This occurs because the experimental treatment is generally not yet available and is being considered for use at an earlier stage of the pathway. Similarly, switches onto other experimental treatments are problematic if they do not represent standard treatment pathways.

In this report, we tackle the issues associated with the use of surrogates in more detail. We extend the literature review carried out by Davis et al. (2012) from 2012 to 2016, using similar inclusion and exclusion criteria, we look at recent statistical and economic approaches, we interview experts from regulatory and reimbursement bodies, we explore how academic research deals with surrogacy, and we report on a workshop in which we bring together representatives from these groups along with experts from POI member bodies.

Four main objectives have been achieved during this analysis. First, we aimed at analysing the methods used to extrapolate from PFS to OS in the field of oncology. In this sense, our results coincide with the results of Davis et al. (2012) in that co-integration and linear regression are the main methodologies applied to extrapolate from PFS to OS. Nevertheless, there exists high variation in the characteristics of the models used that hinder the correct comparison of results.

The second objective was to identify whether a clear guidance exists in the literature about what is considered to be 'best practice' in extrapolation from PFS to OS. Although the original criteria suggested by Prentice (1989) are still valid, the literature review indicates that more robustness and standardisation than those originally proposed are needed to correctly test PFS surrogacy. For instance, the work of Buyse et al. (2016),

Buyse et al. (2007) and Buyse et al. (2000) has highlighted the importance of analysing both aggregate clinical trial and IPD data. In addition, their analyses have focussed attention on the concept of Surrogate Threshold Effect (STE). However, the experts interviewed mentioned that the statistical work of Buyse et al. (2016) has had very little application so far, despite being available in earlier forms for over a decade. Additionally, some recent economic research (Stevens et al., 2014) opens a new avenue for determining the cost effectiveness of a surrogate endpoint by weighing up the benefits from earlier adoption against the likelihood that earlier estimation of benefits using a surrogate will be less precise than that of estimation using overall survival. The benefits of earlier adoption have not been routinely included in economic evaluation of surrogate endpoints.

In addition, we aimed at evaluating whether the relationship between PFS and OS vary by tumour type. The cancer that has been studied most frequently in the literature of surrogacy is lung cancer, particularly NSCLC. Colorectal cancer and renal carcinoma are also among the most mentioned in the literature. However, no consistent results have been found by cancer type. This is explained in part by important variations in the statistical estimation methodology used to support the results. These variations are observed between cancer types and within the same cancer type.

A number of factors affect the relationship between PFS and OS. Therefore, our results, as well as those of Davis et al. (2012), indicate that there is no unique correct answer for the question of whether PFS is an appropriate surrogate for OS in oncology. Moreover since we have not conducted a systematic literature review of any particular cancer type, we cannot say whether PFS should or should not be used in any particular case. What this study concludes is that any consideration of evidence relating to PFS should consider both tumour type and other factors such as the line of treatment and the type of therapy.

Many of these factors are related to the length and characteristics of post-progression survival (PPS). This indicates that understanding the factors that drive PPS is crucial for the analysis of whether PFS is a good surrogate for OS. The extent to which protocols of future follow-up of clinical trial patients consider procedures for gathering information of factors that reflect the effect of the post-progression management of the disease, it should be possible to extract stronger conclusions from the statistical analysis to validate PFS as a surrogate of OS. Thus, research in this area is indicated.

Our fourth objective was to determine the key limitations, weaknesses and gaps in the current literature and methods used to test PFS surrogacy, where future research efforts might usefully be targeted. In this regard, we have identified a lack of rigour on the application of the linear regression methodology. It is crucial that authors and policy makers involved in the discussion of the role of PFS as a surrogate for OS know the correct methodology to be applied and validated . A further issue in the analysis of surrogacy is the heterogeneity relating to the definition of progression among clinical trials. This and the difficulty in finding the required information in clinical trial reports indicates a need for standardisation of clinical trial protocols that allows for comparability between trials in the same cancer type.

One issue that came to the fore during this project was that very little research has been undertaken using individual patient data (IPD) from trials. Using only a single estimate for PFS and a single estimate for OS from a reported trial wastes an enormous amount of data. Data have been collected by manufacturers on an individual patient basis, and

much of the richness of these data has not been taken advantage of. If, as seems likely, the use of IPD would allow much greater precision and efficiency in determining where a putative surrogate endpoint does indeed predict the final outcome of interest, it would seem to be in the interests of both manufacturers and society that that data be made more freely available. It should speed up the process of innovation, leading to better and more consistent decision-making and potentially more effective and cost-effective healthcare products per year being put on the market.

Finally, it is important to highlight that PFS is not only important as a surrogate measure for OS, but it is also an important endpoint in itself. PFS is key to understanding the effect of an intervention on the tumour burden process; this is the mechanism through which anticancer agents are expected to provide benefit. This is one of the reasons that explain why USFDA and the EMA consider PFS to be an accepted regulatory endpoint to support cancer drug approval (US Food and Drug Administration (USFDA), 2007a; European Medicines Agency (EMA), 2012).

Abbreviations

ctDNA	Cell-free tumour DNA
DCR	Disease control rate
DFS	Disease-free survival
EMA	European Medicines Agency
HR	Hazard ratio
IPD	Individual patient data
IQWIG	Institute for Quality and Efficiency in Health Care
mRECIST	Modified response evaluation criteria in solid tumours
NICE	National Institute for Health and Care Excellence
NSCLC	Non-small cell lung cancer
OLS	Ordinary least squares
OS	Overall survival
PD	Progressive disease
PFS	Progression-free survival
PFS2	Progression on next-line therapy
PPS	Post-progression survival
PSA	Probabilistic sensitivity analysis
QoL	Quality of life
RCT	Randomised controlled trial
RECIST	Response evaluation criteria in solid tumours
SCLC	Small cell lung cancer
STE	Surrogate threshold effect
TTF	Time to treatment failure
TTP	Time to progression
USFDA	US Food Drug Administration
WHO	World Health Organisation
WLS	Weighted least squares

1. INTRODUCTION

The outcomes from clinical and other healthcare trials of most interest to patients and health systems are usually increases in the length and quality of life as a result of treatment or other intervention. Many reimbursement agencies around the world measure effectiveness in terms of a quality adjusted life year (QALY) gain, which allows treatments in different disease areas to be assessed against a common metric, facilitating efficient resource allocation decision making through economic evaluation. QALY gains represent the quality and length of life benefits associated with new treatments. In many if not a majority of cases, the increase in length of life, that is overall survival (OS), is the most important factor in decisions on uptake, having the largest impact on assessments of clinical and cost-effectiveness.

This poses a problem, because complete information on the increase in length of life associated with a new treatment is not available until the last person in the trial dies. In the vast majority of cases trial data are "censored" (that is, cut short before the last patient dies). For economic evaluation, it is generally accepted that for any intervention that affects survival, an economic model must take a life-time perspective and therefore estimates of the OS advantage associated with a new treatment are required. The most common approach for arriving at these estimates is to use survival analysis methods to extrapolate beyond the observed trial period. Parametric models, which make various assumptions about the underlying hazard and survival functions, are fitted to trial data and are used to extrapolate out into the future, as described by Latimer (2013). However extrapolation of OS data is not a straightforward task – indeed it may be argued that events occurring within the trial may invalidate the observed OS data, making it unsuitable to extrapolate from. For instance, often progression-free survival (PFS) is used as the primary endpoint in oncology trials and once this event has been observed, treatment switching is permitted. If the treatments switched to do not reflect standard care pathways this could mean that it is not appropriate to extrapolate out from the observed OS data. Methods to adjust for treatment switching have been suggested and used (Latimer et al., 2014) but whilst useful, these methods are imperfect. An alternative to extrapolating from observed OS data is to use PFS as a "surrogate" outcome. This report focuses on the use of surrogate endpoints.

A surrogate endpoint in a trial is an endpoint that is related to a patient-relevant endpoint that may not be available for many more years, such as overall survival (OS) (which only becomes available for an individual after his or her death).

Surrogate endpoints are used as primary endpoints in some trials so that they can be shorter and involve fewer patients, and thus bring a new drug or treatment to the market sooner. This should allow the benefits of successful treatments to be realised at an earlier date; this is not only beneficial to patients and health systems, but will provide greater returns to manufacturers. Products which would not have been worthwhile for manufacturers to produce without the use of a surrogate outcome will become so if the returns on earlier adoption using a surrogate are sufficiently high.

However, these benefits are achieved at the expense of relying on a less accurate measure of final outcome than would occur by waiting for the requisite data, such as OS, to become available. Thus there is a trade-off between time elapsed from the end of study before information becomes available and the accuracy of the information about the benefits of treatment. If the surrogate outcome is not a sufficiently accurate

predictor of final outcome, the use of the surrogate endpoint will not necessarily yield a net benefit to society.

It is necessary, therefore, to determine what characteristics a surrogate outcome should have that are likely to make it a better alternative to waiting until there is enough accuracy in the measurement of the final outcome. It is widely known that the appropriateness of a surrogate will depend on the nature of the disease (type and stage of cancer), the nature of the treatment, the trial methods, and the characteristics of the patients. It is also recognised that a viable surrogate variable must be highly correlated with the 'true' endpoint variable. However, what appears to be less well-known is that this is not sufficient, because the usefulness of a surrogate also depends on whether the treatment under consideration affects the surrogate endpoint in a similar fashion to that of the true endpoint.

Two further problems in oncology with surrogate endpoints concern earlier-line therapy, given that further therapy and/or subsequent diagnosis will usually take place before death. With more than one intervention in a treatment pathway, any attempt to attribute OS to one of the treatments (or diagnosis) rather than another is virtually impossible. As previously stated, such post-study treatments may lead an analyst to conclude that it would be preferable to estimate OS based upon measurements of PFS as a surrogate, rather than by extrapolating observed OS data. However, these post-study treatments also cause problems for the use of surrogates, because they may make it very difficult to ascertain the true relationship between the surrogate (PFS) and the desired outcome (OS).

In cases where numerous post-study treatments are given, PFS and OS are likely to have a low level of correlation. This should not automatically be considered a problem, since the decision problem being addressed is important to take into account. Typically, we are interested in the effectiveness and cost-effectiveness associated with inserting a new treatment into the treatment pathway. If post-study treatment given in a trial are with experimental agents that are not part of the standard treatment pathway the distortion that they create in any analysis of OS or the PFS:OS relationship is problematic, because in reality these treatments would not be given. To address the decision problem the observed OS data is not appropriate and therefore the observed PFS:OS relationship is not appropriate. Conversely, if the post-study treatments given are representative of standard treatment pathways, these are reflective of what would be likely to be given in reality. Thus the observed OS data and PFS:OS relationship are valid for the decision problem being addressed. If a multitude of standardly available post-study treatments means that PFS benefits are not translated into OS benefits, it is appropriate for this to be reflected in an economic evaluation undertaken to inform resource allocation decision making.

That said, it may often be the case that post-study treatments are not representative of standard care and therefore the observed OS data are problematic. This is more likely to be the case for earlier line therapy because at this stage the subsequent treatment pathway is likely to be longer. Therefore the use of surrogates may be more important for earlier-line therapy, whilst also being more difficult to validate.

Another reason why this is likely to be true is that death will be further into the future for earlier-line therapies. For later-line therapies, it is more feasible to collect post-trial data on survival and quality of life. This highlights a further problem in oncology and oncology trials: there are currently few incentives or systems for the collection of quality of life (QoL) data after the end of the trial period. Even if retrospective analyses can

compare trial outcomes on surrogates with OS (from registry data), there are no QoL data post-trial (and in some cases, none within the trial).

This project aims at: (1) analysing the methods used to extrapolate from PFS to OS in the field of oncology; (2) identifying whether a clear guidance exists in the literature about what is considered to be 'best practice' in extrapolation from OS to PFS; (3) evaluating whether the relationship between PFS and OS varies by tumour type; and (4) determining the key limitations, weaknesses and gaps in the current literature and methods used to test PFS surrogacy, where future research efforts might usefully be targeted.

In order to do so the study was divided into three the stages: (1) a review of the clinical evidence, updating an earlier study by Davis et al. (2012); (2) interviews and discussions with regulators, the "reimbursement" body in England (National Institute for Health and Care Excellence, NICE) and researchers into surrogate endpoints; (3) a workshop that included representatives of all of the above groups, as well as a NICE Appraisal Committee Chair and representation from six of the POI member bodies and from OHE.

Section 2 sets out the methodology and the results extracted from the literature review. Section 3 describes the interviews and workshop and the work of leading methodologists in statistical and modelling analyses of surrogate endpoints. Finally, section 4 concludes by summarising the results and discussing some policy implications.

2. LITERATURE REVIEW

We conducted a review of the literature, to identify the 'state of the art' in extrapolating from clinical endpoints to overall survival with a particular focus given to PFS in oncology trials. Previous literature has been conducted on this topic by Davis et al. (2012). This work summarised the methodologies applied and the challenges faced in the extrapolation of PFS to OS. They identified 266 articles, using citation searching (conducted in Medline and the Science Citation Index) to identify relevant papers from an initial list of three papers already known to the authors. The authors state that a systematic literature review was not feasible, first, because an exploratory search returned a very large number of references (over 3,000), and second, because any attempt to make the search more specific resulted in many relevant papers being excluded.

In order to capture those articles in which PFS is a relevant measure, Davis et al. (2012) included any form of cancer in which the treatment intent is palliative rather than curative. In addition, they included all reviews that examined the statistical relationship between OS and either PFS or time tom progression (TTP). Papers simply reporting the target outcomes from single trials or multiple trials were not included. They identified 19 key articles concerning the relationship between PFS/TTP and OS in advanced or metastatic cancer (Table 1) and extracted the main results from them. In general, the Davis et al. (2012) review suggests that the evidence collected until 2012 supports a significant relationship between PFS/TTP and OS. However, their results indicate that the strength of this relationship varies extensively between cancer type and within cancer type. They attribute this variation to the dissimilarities in studies' characteristics, such as tumour type, the line of therapy, and the diversity of methods used.

Table 1. Articles identified by Davis et al. (2012) and number of citations between 2012 and 2016

Article	Citations *	Cancer type	Type of article	Number of Clinical Trials	Number of Patients	R^2	Correlation
Louvet et al. (2001)	75	Colorectal	Aggregate clinical trial data	29	13,498		0.481
Hackshaw et al. (2005)	44	Breast cancer	Aggregate clinical trial data	42	9,163	0.56	
Johnson et al. (2006)	121	Colorectal and NSCLC	Aggregate clinical trial data	Colorectal 146 / NSCLC 191	Colorectal 35,557 / NSCLC 44,125	Colorectal 0.33 / NSCLC 0.19	
Buyse et al. (2007)	231	Colorectal	IPD and Aggregate clinical trial data	10	3,089	0.98	0.82 / ‡
Tang et al. (2007)	169	Colorectal	Aggregate clinical trial data	39	18,668	0.65	Median 0.79 / Differences in median 0.74
Bowater, Bridge and Lilford (2008)	15	Breast, colorectal, hormone refractory prostate and NSCLC	Aggregate clinical trial data	Breast 33, colorectal 38, refractory prostate 23 NSCLC 13	NA		Not significant
Burzykowski et al. (2008)	206	Breast cancer	IPD and Aggregate clinical trial data	11	3,953		Individual 0.688 / Trial level 0.48
Miksad et al. (2008)	57	Breast cancer	Aggregate clinical trial data	31	4,323	Anthracyclines 0.49 / Taxanes 0.35	Kappa test: Anthracycline 0.71/ Taxanes 0.75
Sherrill et al. (2008)	47	Breast cancer	Aggregate clinical trial data	67	17,081	0.30	Kappa test 0.47
Halabi et al. (2009)	79	Prostate	IPD	9	1,296		0.3 / ‡
Hotta et al. (2009)	35	NSCLC	Aggregate clinical trial data	54	23,457	Univariate 0.33/ Multivariate 0.41	
Polley et al. (2009)	38	Brain	IPD	3	193		‡
Wilkerson and Fojo (2009)	37	No particular cancer type	Aggregate clinical trial data	66	NA	Differences in median 0.49/ HRs 0.62	
Mandrekar et al. (2010)	26	NSCLC	IPD	4	284		‡
Bowater, Lilford and Lilford (2011)	9	Breast cancer and colorectal	Aggregate clinical trial data	Breast 95/ Colorectal 74	NA	Breast 0.37 / Colorectal 0.11	
Foster et al. (2011)	41	SCLC	IPD and Aggregate clinical trial data	9	870	Trial level 0.79	Individual level: ‡ Trial level: 0.75 / 0.80
Heng et al. (2011)	52	Renal cell carcinoma	IPD	NA	1,158		Kendall's tau 0.42/ Fleischer model 0.66/ ‡

Hotta et al. (2011)	37	NSCLC	Aggregate clinical trial data	70	38,721	0.256	
Chirila et al. (2012)	13	Colorectal	Aggregate clinical trial data	62	23,527	0.48	Pearson 0.89/ Spearman 0.78
Total citations	**1,332**						
After removing duplicates	**790**						

* Results of the authors' search of the number of citations of each article between 2012 and 2016
‡Landmark analysis
Source: Data extracted from Table 1 and Table 2 of Davis et al. (2012)

We updated the evidence found by Davis et al. (2012) to the present (2016) using the same scoping techniques to preserve as much continuity as possible between the earlier (Davis et al., 2012) work and our more recent work. Specifically, we sought to document advances in the topic. A citation search from January 2012 to June 2016, using Google Scholar, identified a total of 790 articles which had cited those original 19 papers.

A previous follow-up of Davis et al. (2012) was carried out by Ciani et al. (2014). Our analysis differs from that of Ciani et al. (2014) as our aim is to examine the statistical methodologies most commonly applied as well as the main limitations that authors face when attempting to correctly assess PFS validity.

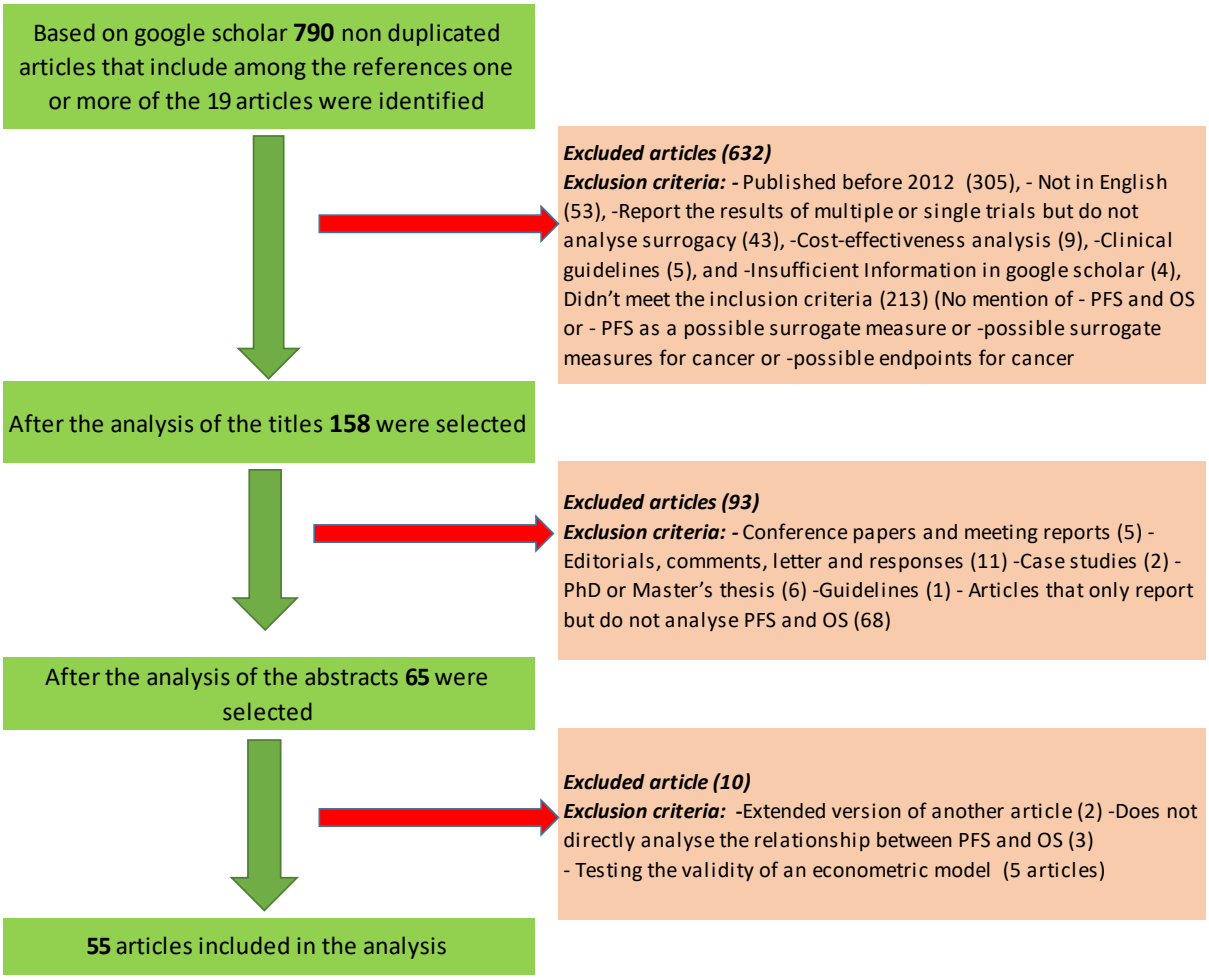
At the end of the review, we report in a further subsection two important recent papers that were not captured by our literature review that consider the problem from the varying viewpoints of the statistician and the economic modeller.

By looking at the titles of the articles, we excluded from the main search those articles not published in English and those that reported multiple or single clinical trials with mention of a surrogate endpoint somewhere in the paper but nothing more than that; cost-effectiveness analyses; or clinical guidelines. We applied four inclusion criteria: 1) articles that mentioned progression-free survival and overall survival in the title; 2) articles that mentioned progression-free survival as a surrogate (including the terms 'surrogate outcome' or 'surrogate endpoint' and 'surrogate measure'); 3) articles that analysed possible surrogate measures for cancer; 4) and articles that analysed endpoints for cancer (Figure 1).

We reviewed the abstracts of 159 articles that fulfilled the above criteria and excluded conference papers, meeting reports, editorials, comments, letters, responses, case studies and PhD/Master theses. During the abstract analysis, we applied one additional selection criterion: we included only those articles that analysed Progression-free Survival as a surrogate marker for Overall Survival, and did not simply report PFS and OS data. 65 documents were fully read, resulting in 55 articles being included in the analysis (Figure 1).

Information regarding methodology, data and factors affecting the relationship between PFS and OS was extracted. Information related to author affiliation and publication journal was also collected. The information was collected by one of the authors (KHV) and the quality and accuracy of the extraction was verified by a second author (AF).

Figure 1. Literature search and selection of articles



Source: Authors' search

1.1. Current approaches and challenges to analyse PFS as a surrogate of OS

Regarding what is considered to be 'best practice' in extrapolation from OS to PFS, the most-quoted work is the analysis done by Prentice (1989). He was one of the first to establish the criteria that an adequate surrogate needs to fulfil in order to demonstrate that a potential measure can be a good surrogate for OS: (1) the surrogate endpoint should predict the clinical endpoint, and (2) the effect of a treatment on the surrogate endpoint should predict the effect of that treatment on the true endpoint.

Although the original criteria suggested by Prentice (1989) are still valid, the literature review indicates that more robustness and standardization than those originally proposed are needed to correctly test these criteria. In this regard, the work of Ciani et al. (2014) summarised three different frameworks that are currently applied to validate

the strength of the evidence: (1) German Institute of Quality and Efficiency in Health Care (IQWiG) framework, (2) Biomarker-Surrogacy Evaluation Schema (BSES3) framework and (3) Elston and Taylor's framework. All of them include the original criteria of Prentice (1989), but also analysed factors that could influence the strength of the relationship, such as the quality of the data and the characteristics of the clinical trial included. In addition, Buyse et al. (2000) proposed that in order to validate a surrogate endpoint it is necessary to analyse both individual level and trial level data. This because the analysis of clinical trial information is particularly important for testing the relationship between the treatment effect on PFS and the treatment effect on OS while the IPD allows the analysis of the relationship between the absolute value of PFS and the absolute value of OS. In order to verify surrogacy, both the relationship between the treatment effects and the relationship between the absolute values should be significant.

In addition, Ciani et al. (2016) state that surrogate endpoints need to satisfy three conditions. First, the level of evidence supporting the relationship between the surrogate endpoint and the desired outcome needs to be considered. A strong correlation should be observed between the surrogate and the end point based on individual patient data as well as between the treatment effect on the surrogate end point and the final outcomes across multiple randomised trials. Second, the strength of the association between the surrogate endpoint and the final outcome should be measured throughout approaches such as regression and meta-analysis. Third, it is necessary that the effect on the final outcome could be predicted and quantified based on the effect on the surrogate, using methodologies such as the surrogate threshold effect (STE) (see section 1.1.2). The effect of the treatment on PFS must be large enough to predict an improvement in OS. This is key for decisions on coverage and reimbursement, because while regulators assess drugs based on safety and efficacy, reimbursement bodies must take into account an intervention's benefit compared with other treatments, as well as its cost.

Table 2 shows the selected articles classified according to type of article. Four types of article were identified. The first three categories refer to articles that directly estimate the relationship between PFS and OS using IPD and/or clinical trial information. 48 out of the 55 articles are classified in these three categories. The fourth category corresponds to a group of articles that summarise previous studies.

Twelve of the 19 articles (63%) identified by Davis et al. (2012) used aggregate data from multiple trials (Table 1). In our literature review, 32 out of the 48 articles use aggregate data from multiple trials which correspond to 67% of the sample. We identified six articles (12.5%) in which both clinical trial data and IPD were used (Table 2) to estimate the relationship between PFS and OS; Davis et al. (2012) identified three (15.8%) (Table 1). This suggests there is still a lack of available and comparable information that hinders the use of both clinical trials and IPD during the validation of a surrogate endpoint. Only a small number of researchers have access to the necessary information that allows them to fulfil the criteria suggested by Buyse et al. (2000). The lack of IPD information is evident if we consider that only 21% of the articles found by Davis et al. (2012) and 21% of the articles found here are based on IPD data. Moreover, six out of the ten IPD articles are based on information collected in a single Japanese institution, which indicates that the extrapolation of the conclusions to a different context should be undertaken with caution (Imai et al., 2014; Imai et al., 2015; Kasahara et al., 2015; Yoshino et al., 2014; Yoshino et al., 2015).

In addition to the 48 articles, we identified seven articles that correspond to a summary of previous studies. Three of the seven articles reported a systematic review of the literature (Ciani et al., 2014; Prasad et al., 2015; Kim and Prasad, 2016) while the other four analysed the relationship between PFS and OS based on previous literature, but without stating how the studies were identified (Table 2).

Regarding the general conclusions, 53% of the articles that used aggregate data from multiple trials support PFS as an appropriate surrogate for OS. Ten of the remaining 15 articles suggest that the validity of the surrogacy depends on factors such as the line of therapy (first vs second or third) (Adunlin, Cyrus and Dranitsaris, 2015; Özer-Stillman et al., 2015) or the type of treatment (Johnson, Liauw and Lassere, 2015). Among the articles that used IPD, the analysis suggests a similar proportion of articles support surrogacy (4/10) as refute it (5/10) (Table 2).

Table 2. General conclusion regarding the relationship between PFS and OS by type of article

Type of Study	Appropriate surrogate	Not an appropriate surrogate	Depends on particular factors	No recommendation
Individual data	Galsky et al. (2013); Halabi et al. (2014); Négrier et al. (2014); Shitara et al. (2013)	Imai et al. (2014); Imai et al. (2015); Kasahara et al. (2015); Yoshino et al. (2014); Yoshino et al. (2015)	Laporte et al. (2013)	
Aggregate data from multiple trials	Beauchemin et al. (2014); Bria et al. (2015); Cartier et al. (2015);Chen et al. (2015); Delea et al. (2012); Félix et al. (2013); Flaherty et al. (2014); Giessen et al. (2013); Giessen et al. (2015); Han et al. (2013); Li et al. (2012); Petrelli and Barni (2013a); Petrelli and Barni (2013b); Petrelli et al. (2015); Shitara et al. (2012); Sidhu, Rong and Dahlberg (2013); Singh, Wang and Law (2014)	Aboshi, Kaneko and Narukawa (2014); Ciani et al. (2015); Petrelli and Barni (2013c); Shitara et al. (2014); Terashima et al. (2015);	Adunlin et al. (2015); Amir et al. (2012); Hotta et al. (2015); Hotta et al. (2013); Johnson et al. (2015); Kawakami et al. (2013); Moriwaki et al. (2016); Özer-Stillman et al. (2015); Petrelli and Barni (2014); Suzuki et al. (2015)	
Individual and aggregated trial data	Agarwal et al. (2014); Shi et al. (2015); Foster et al. (2015); Mauguen et al. (2013)	Michiels et al. (2016); Paoletti et al. (2013)		
Review of previous studies (OS vs PFS relationship)		Prasad et al. (2015)	Matulonis et al. (2015)	Ciani et al. (2014); Garon (2012); Kim and Prasad (2016); Pilz and Manegold (2013); Sherrill et al. (2012)

Source: Authors' analysis

For the articles based on aggregate data, Davis et al. (2012) found that the number of trials included ranged between 13 and 191 (Table 1). As shown in Table 3, the maximum number of trials included has decreased in comparison with the results presented by Davis et al. (2012). However, a high variation still exists within the number of articles considered.

The number of observations analysed is highly important for robust conclusions in quantitative analysis. In this regard, in most of the selected articles, some clinical trials identified by the researchers reported more than one treatment arm (27/32). This has the positive effect of considering a higher number of observations in the analysis which may result in more robust conclusions. Among these articles, the number of different treatment arms included in the analysis ranges from 7 to 230. However, to include multiple treatment arms has the disadvantage that some observations are naturally grouped. This could lead to problems in the quantitative analysis such as heteroscedasticity.[1] Heteroscedasticity makes it difficult to test whether PFS is a significant variable for explaining OS in the results from a linear regression model, although it does not affect the R^2 value of the regression. In the majority of these articles, each treatment arm is considered as a separate observation. However, there are examples of where the authors based the analysis on only one treatment arm per clinical trial or the data of the different experimental arms are combined (e.g. Paoletti et al., 2013).

Table 3. Number of patients and trials included in the selected articles

	Number of patients			Number of trials		
	Minimum	**Average**	**Maximum**	**Minimum**	**Average**	**Maximum**
Individual data	35	537	2,331	1	4	7
Aggregate data from multiple trials	2,148	13,031	43,459	6	44	153
Individual and aggregated trial data	689	5,238	16,762	8	17	29

Source: Authors' research

1.1.1. Cancer Types

Like the conclusions of Davis et al. (2012), the results from the validation of PFS as a surrogate point for OS varies substantially among cancer types and there is little or no consistency in the results by cancer type. In Table 4, lung cancer, with 16 of the articles (30%), dominates the discussion: 11 related to non-small cell lung cancer (NSCLC), three to small-lung cell cancer and two that considered all lung cancer types. Although a number of articles (6 out of 16) suggest that PFS is not an appropriate surrogate for OS, there is no consistency in the results even when we consider the line of therapy and the phase of the clinical trial.

The second most mentioned cancer type is colorectal cancer (Table 4). Here the probability of concluding that PFS is an appropriate surrogate point for OS is higher (5 out of 7), which is in agreement with previous studies that lead to the common understanding that colorectal cancer is one of the few cancer types for which there is agreement in the validity of PFS as a surrogate of OS. Following colorectal cancer in the frequency of papers is renal carcinoma where four articles analysed first line therapy and two articles first line or beyond. Four out of the five articles analysed more than one

[1] Heteroscedasticity refers to the case in which the variability of a variable varies across the range of values of a second variable that predicts it.

cancer type, corresponding to articles classified as reviews of previous studies (Table 2). Only one article that analysed more than one cancer type is based on aggregate clinical trial information: Amir et al. (2012). They use data from 26 randomised trials conducted between 2002 and 2012 that have led to drug approval by the US Food and Drug Administration.

Buyse et al. (2000) argue for greater confidence in the relationship between surrogate and OS when correlation is established across studies with different characteristics. Therefore, different factors that can affect the link between PFS and OS have been analysed in the literature. A deeper analysis of these factors is undertaken in section 1.1.4. Here, we introduce the factor that appears to be particularly important for verifying the PFS and OS relationship: the line of therapy. In those cases where the analysis cannot validate the surrogacy for first line therapy, as distinct from the second or third line therapy, the first possibility considered by the researchers is that the effect of the treatments that affect post-progression survival is higher and so obscures the results of the first line treatments. The importance of validating the surrogacy in the first line is observed in Table 4 where most of the articles include data of first line therapy. However, we were not able to find any pattern related to the therapy line in terms of the general conclusion of the articles. This means that there is no particular therapy line for which there is a higher probability that the analysis validates the surrogacy of PFS, which suggests that this is not the only characteristic that affects the results. In section 1.1.4 we will discuss in detail some the factors that the literature suggests are determinants of the relationship between PFS and OS.

Table 4. Number of articles per cancer type

		General conclusion				Line						Phase				
	Total	Appropriate surrogate	Depends on particular factors	No appropriate surrogate	No recom.	1st	2nd	1st and 2nd	1st and beyond	2nd and beyond	NA	II	III	I, II and III	II and III	NA
Non-small cell lung cancer	11	2	4	4	1	4	1	1	3	1	1	1	3	4		3
Colorectal cancer	7	5	1	1		3	1		3				2	1	3	1
Renal cell carcinoma	6	4	1	1		4		1	1				4	2		
Gastric cancer	5	2	1	2		3	1		1					3	1	1
Breast cancer	3	1	1	1		1			2				1	1	1	
Small cell lung cancer	3	1		2		3								1		2
Lung cancer	2	1			1				1		1					2
Multiple myeloma	2	2							1		1				1	1
Urothelial carcinoma	2	2				1	1							2		
Biliary tract cancer	1		1			1								1		
Gastrointestinal stromal tumor	1		1						1					1		
Glioblastoma	1	1							1					1		
Hepatocellular carcinoma	1			1		1										1
Melanoma	1	1				1								1		
Nasopharyngeal carcinoma	1	1									1				1	
Neuroendocrine	1	1									1				1	
Ovarian Cancer	1		1								1					1
Pancreatic cancer	1	1				1								1		
No particular cancer type	5		1	1	3				3		2		1		1	3
Total	**55**	**25**	**12**	**13**	**5**	**23**	**4**	**2**	**17**	**1**	**8**	**1**	**11**	**19**	**9**	**15**

Source: Authors' research

1.1.2. Methodologies and statistical results

Table 5, Table 6 and Table 7 show the methodologies applied and types of estimation undertaken by the authors of the selected articles. The most usual preference is for the use of correlation (particularly Spearman, Pearson and Kendall's τ) (34/48) and weighted or unweighted linear regression (35/48) to validate PFS as a possible surrogate of OS, which is comparable to the findings of Davis et al. (2012). Moreover, like Davis et al. (2012), our results suggest that although most of the selected articles use a correlation or linear regression model, many different variations in methodology apply which make it difficult to compare the results. For example, Aboshi et al. (2014) use uni- and multi-variate logistic regression analyses to estimate the relationship among PFS and OS where weights relating to the number of observations were not applied. In addition, Bria et al. (2015) estimate the correlation of the probability of PFS at 3 and 6 months and OS at 9 and 12 months (Table 5).

However, beyond the variations that each author applies in the analysis, the main question to be answered is whether the recent investigations in the topic include an adequate methodology and use suitable data to test PFS as a possible surrogate for OS. The common practice for testing whether the surrogate is capable of predicting the clinical end point is to analyse the relationship between the absolute values of the PFS and OS. Those articles that include aggregate clinical trial information mostly use median PFS and median OS, but in some cases the estimation uses a logarithmic transformation of the variables (Giessen et al., 2015). In the case of IPD articles, the relationship between the absolute values per patient of PFS and OS is used to prove the predictive capacity of PFS. A sub-group of the IPD articles uses a landmark approach where the PFS rate at different months (3, 6, 9 and/or 12) in predicting OS is analysed (Table 6 and Table 7).

The effect of treatment in changing PFS to predict the effect of treatment in changing OS is explored throughout the analysis of aggregate clinical trial information. Here it is common to compare the hazard ratio (HR) of the two endpoints; however, we also identified articles in which the differences in median PFS and OS are examined. In addition, a variety of methodologies have been followed to overcome the problem of not reporting HR.

Regarding the results by cancer type, NSCLC, which has the highest number of observations, in most of the cases shows statistical (R^2 and correlation) results lower than 0.7, which indicates a weak validity of PFS as a surrogate measure for OS. Three particular exceptions should be mentioned. First, the study of Suzuki et al. (2015) (Table 5), with a Spearman correlation value equal to 0.77, supports only PFS surrogacy if PPS length is less than 6 months. In addition, Imai et al. (2014) and Yoshino et al. (2015) estimate Spearman correlation values of 0.76 and 0.72 respectively. These can be considered high in comparison with other studies; however, they concluded that PFS is not an appropriate surrogate for OS based on the low values of R^2.

Among all the results, Petrelli and Barni (2013b) stand out because of the lack of consistency between supporting PFS surrogacy and having low correlation and low R^2 values. They analysed the surrogacy of PFS in the context of NSCLC and observed weak correlation between PFS and OS. However, they still support PFS as a surrogate for OS, a decision that appears to be influenced by the slope of the linear regression. The slope suggests that a one-month gain in PFS will be linked to three weeks' prolongation in OS. However, in view of the very small R^2 value, the reasons for conclusion are not clear.

In addition to NSCLC, Table 6 and Table 7 show four more examples that refer to lung cancer (three for small cell lung cancer and one for lung cancer in general). Two of the four articles are based only on IPD, and none of them supports PFS as a surrogate (Table 6 and Table 7). In addition, Foster et al. (2015) used both IPD and aggregate clinical trial information. The results based on clinical trial information led the authors to support the validity of the surrogacy, while IPD data are less conclusive (Table 7). The remaining study, Mauguen et al. (2013) also used IPD and clinical trial data, but, contrary to Foster et al. (2015) where only the clinical trial information fully supported surrogacy, IPD also supports PFS as an appropriate surrogate for lung cancer (R^2 and correlation higher than 0.7) (Table 7).

Table 5. Articles based on aggregate clinical trial information: Methodology and estimations

Author	Cancer Type	#Patient (P) / #Clinical Trials (CT)	Correlation	Treatment effect corr.	Absolute value corr.	Weighted Linear Regression	Treatment effect R^2	Absolute value effect R^2	Characteristics of the methodology
Appropriate surrogate									
Delea et al. (2012)	Renal cell carcinoma	P 10,943/ CT 31	Pearson/ Spearman	lnHRs: 0.80/ Difference in median: 0.54		yes	lnHRs: 0.63/ Difference in median: 0.28		1) For studies that did not report HRs for PFS/TTP or OS, HRs were estimated using data from Kaplan–Meier curves or numbers of events and log-rank statistics 2) median survival was estimated by fitting Weibull survival functions to reported Kaplan–Meier curves
Li et al. (2012)	NSCLC	P 9,903/ CT 60				yes		Simple regression: 0.70/ Multivariate regression: 0.74	1) Multivariate linear regression models 2) Significance of surrogate as a survival marker compared: Area under their receiver operating characteristic (ROC) curves. 3) Discrimination ability tested.
Shitara et al. (2012)	Gastric cancer	P 10,484/ CT 36	Spearman	0.80	0.70				
Félix et al. (2013)	Multiple myeloma	P 22,696/ CT 153	Spearman		0.75				1) Two-step approach to a simultaneous Tobit model. Censored normal-weighted regression with the robust option in Step 2. 2) Heteroscedasticity: Generalized Method of Moments (GMM).
Giessen et al. (2013)	Colorectal cancer	P 22,736/ CT 50				yes	0.87	0.86	
Han et al. (2013)	Glioblastoma	P 7,125/ CT 91	Pearson			yes	0.92	0.7	1) Lead-time that could be gained by using PFS instead of OS
Petrelli and Barni (2013a)	Colorectal cancer	P 16,408/ CT 34	Spearman	0.59	0.64	yes		0.43	
Petrelli and Barni (2013b)	NSCLC	P 4,176/ CT 10	Spearman	0.64	0.26	yes		0.00007	
Sidhu et al. (2013)	Colorectal cancer	P 20,438/ CT 24		0.86		yes	0.73		
Beauchemin et al. (2014)	Breast Cancer	P 43,459/ CT 144	Pearson/ Spearman	0.427	0.428	yes	0.86		1) Linear regression: only studies with statistically significant difference in both PFS/TTP and OS between treatment arms
Flaherty et al. (2014)	Melanoma	P 4,416/ CT 12		0.85					1) Fixed and random effects
Singh et al. (2014)	Neuroendocrine	P 2,584/ CT 22				yes	0.03	0.216	1) Information regarding clinician perceptions of disease progression was extracted from the clinicians survey
Bria et al. (2015)	Renal cell carcinoma	P 8,791/ CT 19	Pearson/ Spearman/ Kendall's τ	0.45 (not sig)	Targeted treatments: 0.85 / Immunotherapy: 0.84	yes	0.66 (not sig)	Targeted treatments: 0.73 / Immunotherapy: 0.71	1) Monthly PFS and OS rates from month 1 to month 12 extracted from publications or Kaplan–Meier curves 2) Cumulative monthly PFS and OS rates: weighted-average approach 3) Treatment arms were merged according to 3 groups: (1) targeted agents, (2) immunotherapy, and (3) placebo. 4) Correlation analysis between 3-month PFS and 9-month OS or 6-month PFS and 12-month OS
Cartier et al. (2015)	Multiple myeloma	P NA / CT 21	Pearson	HRs: 0.82/ logHRS: 0.80		yes	HRs: 0.67 / logHRs: 0.63		

Chen et al. (2015)	Nasopharyngeal carcinoma	P 5,212/ CT 21				yes	0.9		1) Errors-in-variables linear regression model with a conservative reliability coefficient of 0.9 (weighted by trial size) 2) When not available FFS and/or PFS, and OS were determined for treatment arms using published data or survival curves
Giessen et al. (2015)	Colorectal cancer	P 10,800/ CT 23	Pearson		0.73				
Petrelli et al. (2015)	Pancreatic cancer	P 8,467/ CT 30	Spearman	0.78	0.75	yes	0.69	0.6	
	Depends on particular factors								
Amir et al. (2012)	No a particular cancer type	P NA/CT 26		PPS<12 : 0.64 / PPS>12 0.38		yes			**Not an appropriate surrogate when post-progression is long*** 1) Identification of a cut-off for median SPP with the greatest discrimination: non-parametric (spline) smooth function applied to the correlation between the ratio of OS to PFS and to PPS
Hotta et al. (2013)	NSCLC	P 24,158/ CT 34				yes	0.69		**Appropriate surrogates when post-study treatments were seldom employed***
Kawakami et al. (2013)	Gastric cancer	P 11,802/ CT 43	Spearman	0.547	ALL: 0.496/ 2005 of older trials: 0.689 / recent trial 0.282				**Correlation decrease over the years*** 1) Dichotomised (older trials and recent trials, based on 2006): to assess the effect of year of completion of trial enrolment.
Petrelli and Barni (2014)	Breast cancer	P 10,138/ CT 20	Spearman	0.78	0.81	yes	0.73	0.61	**Not appropriate surrogate HER2-negative disease trials, where the weight of PPS is stronger***
Adunlin et al. (2015)	Colorectal cancer	P 0/CT 72	Spearman	HRs: 0.46 / Difference in median: 0.52		yes	HRs: 0.31 / Difference in median: 0.44		**Appropriate surrogate in 2nd line and beyond*** 1) Multivariate regression analysis
Hotta et al. (2015)	NSCLC	P 7,633/ CT 18				yes	0.23		**Molecularly selected patient trials vs all-comer trials: these molecularly targeted trials using PFS would be considered positive if their HR is less than or equal to 0.6 for PFS*** 1) Influence of trial design (molecularly selected patients vs. all-comers) evaluated: multiple stepwise regression analysis 2) Receiver operating characteristic (ROC) analysis was used to identify the most accurate discrimination thresholds
Johnson et al. (2015)	Renal cell carcinoma	P 10,797/ CT 30				yes		EIV regression: 0.49 / OLS: 0.44	**PFS surrogacy is not generalizable across all drug classes*** 1) Errors-in-variables (EIV) regression 2) Evaluated the effect of prognostic covariates
Özer-Stillman et al. (2015)	Gastrointestinal stromal tumor	P 2,189/ CT 14	Pearson		0.72	yes		0.52	**Not appropriate surrogate in first line therapy*** 1) Quality of the evidence assessed: GRADE approach
Suzuki et al. (2015)	NSCLC	P NA/ CT 32	Spearman	0.77					**Appropriate surrogate only with a PPS of less than 6 months*** 1) Optimal point of correlation of PFS-HR and OS-HR by every 1 month of SPP: by using a linear regression model
Moriwaki et al. (2016)	Biliary tract cancer	P 2,148/ CT 17				yes		All: 0.66 / Target therapy: 0.78 / gemcitabine-therapies: 0.78	**PFS is an appropriate end point in a phase II trial of a newly developed drug*** 1) Trials with gemcitabine-containing therapies and with targeted agents analysed separately
	Not an appropriate surrogate								

Petrelli and Barni (2013c)	Renal cell carcinoma	P 3,188/ CT 6	Pearson / Spearman	0.36	0.869	yes	0.07	0.97	
Aboshi et al. (2014)	NSCLC	P 23,337/ CT 65	Spearman		0.689			0.439	1) Covariates analysed by univariate logistic regression analysis using a fixed-effect model and multivariate logistic regression analysis
Shitara et al. (2014)	Gastric cancer	P 4,286/ CT 64	Spearman	0.36	0.56				
Ciani et al. (2015)	Colorectal cancer	P 40,243/ CT 101	Spearman	HRs: 0.75 / Difference in median: 0.59		yes	HRs: 0.34 / Difference in median: 0.52		1) der Simonian and Laird random-effects univariate meta-analyses were used to calculate the pooled treatment effect (95% CI) for OS. 2) Random-effects multivariate meta-analyses.
Terashima et al. (2015)	Hepatocellular carcinoma	P 5,803/ CT 56				yes		0.546	

*Factor which determines the validity of the PFS surrogacy

Source: Authors' research

Table 6. Articles based on individual patient information: Methodology and types of estimation

Author	Cancer type	#Patient (P) / #Clinical Trials (CT)	Correlation	Treatment effect correlation	Absolute value correlation	Linear Regression	Treatment effect R^2	Absolute value effect R^2	Characteristics of the methodology
Appropriate surrogate									
Galsky et al. (2013)	Urothelial carcinoma	P 364/ CT 7	Fleischer		0.86				1) The method of Kaplan and Meier was used to estimate the OS of patients stratified by disease progression at 6 or 9 months. 2) Correlation estimated using the statistical model for dependence between PFS and OS developed by Fleischer, Gaschler-Markefski and Bluhmki (2009)
Shitara et al. (2013)	Gastric cancer	P 291/ CT NA	Spearman		0.75				1) Prognostic factors for PPS: uni- and multivariate analyses using a Cox proportional hazards model.
Halabi et al. (2014)	Renal cell carcinoma	P 1,381/ CT 2	Kendall's τ		0.53				1) The Kaplan-Meier product-limit method was used to estimate the OS distribution by the PFS rate at 3 months and at 6 months 2) The Cox proportional hazards model was used to assess the significance of the effect of the PFS rate at 3 months and at 6 months in predicting OS 3) Multivariable Proportional Hazards Models 4) Adjusted association: 3.145 (considered the fact that OS is always higher than PFS)
Négrier et al. (2014)	Renal cell carcinoma	P 750/ CT 1							1) Weibull parametric model to failure time data was fitted to determine whether longer PFS was significantly and meaningfully predictive of longer PPS. PFS was significantly predictive of longer PPS ($P < 0.001$). 2) In a sensitivity analysis by Kaplan–Meier non-parametric method, PPS curves for three approximately equal numbered groups of patients categorised by PFS were compared by log-rank test.
Depends on particular factors									
Laporte et al. (2013)	NSCLC	P 2,331/ CT 5	Kendall's τ		0.59	yes	0.62-centers 0.72-strata		**PFS surrogacy valid only for treatments that have a major impact on PFS (risk reduction of at least 50%)*** 1) The distributions of PFS and OS were estimated using the Kaplan-Meier method. Treatment groups were compared using a Cox regression model. 2) The association between PFS and OS was quantified through a bivariate copula model fitted on IPD.
Not an appropriate surrogate									
Imai et al. (2014)	NSCLC	P 39/ CT NA	Spearman		0.76	yes		0.50	1) Prognostic factors for PPS: proportional hazards model with a stepwise regression procedure
Yoshino et al. (2014)	NSCLC	P 35/ CT NA	Spearman		0.13	yes		0.45	1) Prognostic factors for PPS: proportional hazards model with a stepwise regression procedure
Imai et al. (2015)	SCLC	P 49/ CT NA	Spearman		0.58	yes		0.24	1) Prognostic factors for PPS: proportional hazards model with a stepwise regression procedure
Kasahara et al. (2015)	SCLC	P 71/ CT NA	Spearman		0.46	yes		0.38	1) Prognostic factors for PPS: proportional hazards model with a stepwise regression procedure
Yoshino et al. (2015)	NSCLC	P 58/ CT NA	Spearman		0.72	yes		0.41	1) Prognostic factors for PPS: proportional hazards model with a stepwise regression procedure

*Factor of which depends the validity of the PFS surrogacy

Source: Authors' research

Table 7. Articles based on both individual patient information and aggregate clinical trial data: Methodology and types of estimation

Author	Cancer Type	#Patient (P) / #Clinical Trials (CT)	Correlation	Treatment effect Correlation	Absolute value correlation	Weighted Linear Regression	Treatment effect R^2	Absolute value effect R^2	Characteristics of the methodology
		Appropriate surrogate							
Mauguen et al. (2013)	Lung cancer	P 5,211/ CT 29				yes	Trial level: range from 0.89 to 0.99	Individual level: range from 0.77 to 0.85	1) Individual level: bivariate survival model that takes censoring into account 2) Trial level: Weighted linear regression model 3) Correlations between 2-year PFS and 5-year overall survival were assessed
Agarwal et al. (2014)	Urothelial carcinoma	P 689/ CT 10	Pearson/ Pearson chi-square - Yates continuity correction	Individual level: 0.45	Trial level: 0.66	yes		Trial level: 0.55	1) Relationship between PFS6/RR and OS12. 2) Estimate of PFS6: generalized linear mixed models with normal random effects for trial 3) Individual level: Pearson chi-square test with Yates continuity correction 4) Trial level: weighted linear regression and Pearson correlation. 5) A second-line phase III trial used for external validation
Foster et al. (2015)	SCLC	P 2,855/ CT 10	Kendall's τ		Individual level: 0.57	yes	Trial level: Copula R^2 - 0.81 / WLS R^2 - 0.77		1) Individual level: bivariate survival model constructed from a Clayton copula with Weibull marginal distributions 2) Trial level: a. weighted (by trial size) least squares regression of marginal Cox model effects (WLS R^2) and weighted (by trial size) correlation of the joint copula effects (copula R^2 and associated standard errors [SE]) 3) Use of data from seven new first-line phase II/III ES-SCLC trials to externally validate findings.
Shi et al. (2015)	Colorectal cancer	P 16,762/ CT 22	Rank Correlation Coefficient		Individual level: 0.51	yes	Individual level: 6 months PFS - 0.69 / 18 months PFS - 0.51 Trial level: WLS R^2 - 0.54 / Copula R^2 - 0.46		1)Individual level: Prognostic value of PFS status at 6 months and at 1 year assessed by the Cox model by using a landmark approach. Correlation: bivariate Copula distribution of the two end points. 2) Trial level: Weighted least squares. Copula R^2 was also estimated.
		Not an appropriate surrogate							
Paoletti et al. (2013)	Gastric cancer	P 4,069/ CT 20	Rank Correlation Coefficient		Individual level: 0.85	yes	Trial level: 0.61		Surrogacy tested within the framework of the GASTRIC meta-analysis
Michiels et al. (2016)	Breast cancer	P 1,839/ CT 8	Spearman		Individual level: 0.67	yes	Trial level: 0.51	Individual level: 0.45	1) Individual level: bivariate survival model that takes censoring into account 2) Trial level: Weighted linear regression model

Source: Authors' research

In relation to colorectal cancer, two out of the five articles that support surrogacy have R^2s and/or correlation values that are lower than 0.7 (Table 5 and Table 7) (Petrelli and Barni, 2013a; Shi et al., 2015). Similarly, Delea et al. (2012) (Table 5) and Halabi et al. (2014) (Table 6) in the context of renal carcinoma support PFS as a surrogate of OS even though the statistical results showed values lower than 0.7. These results and those for lung cancer show a lack of agreement among the authors on the importance assigned to different statistical estimation methods as well as the need for commonly accepted lower limits for correlation and R^2. This is the minimum that should be achieved in order to consider the statistical result as proof of surrogacy. In the case where a surrogate endpoint is strongly believed to lead to substantial cost savings and/or substantial earlier health benefits as a result of a decision for earlier adoption of the treatment in question, this conclusion may need to be reconsidered, because in such cases, a high correlation may not be required.

As Table 5, Table 6 and Table 7 show, one of the two methodologies that dominate the analysis is the weighted linear regression model. This model is based on assumptions that are not tested in the majority of the articles when analysing surrogacy. We identify only a few cases in which the linear regression model assumption was mentioned. An example is the analysis of Félix et al. (2013) where Generalized Method of Moments (GMM) is used to control for heteroscedasticity. In addition, Johnson et al. (2006) mentions that estimation based on immunotherapy clinical trials show unsatisfactory diagnostics with non-normality and heteroscedasticity in the residuals. Li et al. (2012) states that in the diagnostic tests for normality and heteroscedasticity, their estimation is consistent with linear regression assumptions.

In spite of the small number of articles that consider linear regression assumptions, these assumptions are so widely known that when there is no problem it could be considered irrelevant to report them in the analysis. Therefore, we cannot conclude whether linear regression assumptions have been adequately considered in the literature. However, in a number of articles, figures clearly show the presence of outliers. An outlier is an observation that, if dropped from the analysis, changes key estimates notably. Two clear examples are Yoshino et al. (2015) whose analysis is based on IPD, and Moriwaki et al. (2016) who analyse aggregate clinical trial information. Although extreme examples, these cases demonstrate the importance of considering the assumptions of the linear regression model and outliers on the validation of the PFS surrogacy. Out of the 48 studies, only 23% (11/48) consider or mention the problem that the presence of outliers poses for the analysis. In five of these 11 cases, the authors test the sensitivity of the results by applying a "leave-one-out" strategy where each trial is left out once at each step and the surrogate model is rebuilt with the other trials (Chen et al., 2015; Foster et al., 2015; Mauguen et al., 2013; Michiels et al., 2016; Shi et al., 2015). Four of these five studies include both IPD and trial data. Similarly, the six remaining articles test the sensitivity of the results by excluding those trials that are considered outliers (Delea et al., 2012; Félix et al., 2013; Flaherty et al., 2014; Özer-Stillman et al., 2015; Petrelli and Barni, 2014; Singh et al., 2014).

Publication bias is another aspect of the methodology that could have a significant impact on the results, particularly for the aggregate clinical trial articles. Apart from Amir et al. (2012), who analysed randomised controlled trials (RCTs) supporting registration of new anti-cancer drugs approved by the US Food and Drug Administration, all articles using clinical trial data show results of systematic literature reviews. These articles could potentially be limited by publication bias with respect to the articles that are available.

Out of these 31 articles, 11 mentioned publication bias as a possible limitation of the study while six included a step to overcome the possible bias. For example, some articles considered both published and unpublished clinical trials (Hotta et al., 2013; Shitara et al., 2014; Shitara et al., 2012) and others analysed the extent to which bias represents a problem using Egger's regression test (Ciani et al., 2015; Delea et al., 2012; Singh et al., 2014). Nevertheless, nearly half of all articles (14/31) do not even mention publication bias which indicates that the researchers may not have considered this problem during the analysis.

Apart from correlation analysis and the weighted linear regression model, a methodology that has increased in importance is the estimation of the surrogate threshold effect (STE). In Davis et al. (2012), STE was reported in only two out of the 19 papers (Buyse et al., 2007; Johnson et al., 2006), while we identified 11 articles that analyse the STE (Table 8). In general, a STE is understood as the minimum treatment effect that should be observed on PFS in order to predict an OS benefit. This concept has the advantage of not being a yes or no answer to the question of the surrogacy, but a lower bound that if achieved by PFS would indicate that PFS can significantly predict OS. Table 1 shows the results for STE found in the literature review. Once again there is a lack of consistency between the STE results and the main conclusions of the analysis. In addition, five of the six articles that included both IPD and aggregate clinical trial data estimate STE. This concept will be further explored in section 1.3 where we will discuss the work undertaken by Buyse et al. (2016).

Table 8. Surrogate threshold effect (STE)

Author	Cancer type	Type of Study	STE	STE definition
Appropriate surrogate				
Mauguen et al. (2013)	Lung cancer	Both*	From 0.93 to 1.00 depending on therapy	Minimum treatment effect on the surrogate that would be necessary to predict a non-zero effect on OS
Sidhu et al. (2013)	Colorectal cancer	Trial	0.9	Minimum PFS effect that predicts a positive OS effect (i.e., OS Hazard ratio < 1)
Chen et al. (2015)	Nasopharyngeal carcinoma	Trial	PFS vs OS: 0.88 / PFS at 3 years vs 5 years OS: STE = 0.84	Minimum treatment effect on the surrogate necessary to predict an OS benefit
Foster et al. (2015)	SCLC	Both*	0.67	Minimum effect on the surrogate needed to detect a nonzero treatment effect on OS was also calculated: estimated throughout unweighted least squares regression model for the Cox model treatment effects
Shi et al. (2015)	Colorectal cancer	Both*	0.57	Minimum treatment effect on PFS required to predict a nonzero treatment effect on OS
Depends on particular factors				
Laporte et al. (2013)	NSCLC	Individual	0.49-centers 0.53-strata	Minimum treatment effect on PFS required to predict a non-zero treatment effect on OS
Johnson et al. (2015)	Renal cell carcinoma	Trial	All-trials and immunotherapy-only trials failed to demonstrate a STE. A targeted therapy trial needs a PFS difference of at least 3.7 months	The STE is determined as: (i) extract median PFS and OS values for each arm in each trial, (ii) calculate the between-arm difference, (iii) regress the PFS and OS differences, (iv) calculate the 95% prediction limits of the regression, and (v) determine the PFS value where the lower 95% prediction line intersects with the horizontal (PFS) x-axis.
Moriwaki et al. (2016)	Biliary tract cancer	Trial	0.83	Vertical line that transects the upper 95% predictive limit and a median OS ratio equal to 1. Represents the minimum PFS effect to predict a positive OS effect
Not an appropriate surrogate				
Paoletti et al. (2013)	Gastric cancer	Both*	0.56	PFS HR value to predict, with 95% probability, an OS HR less than 1
Ciani et al. (2015)	Colorectal cancer	Trial	0.8	Intercept of the regression line with zero effect on OS
Michiels et al. (2016)	Breast cancer	Both*	0.72	Minimum treatment effect that is necessary on PFS to be able to predict a non-zero effect on OS

* "Both" means aggregate clinical trial information together with IPD.
Source: Authors' research

It is worth citing here the one study that includes IPD and clinical trial information that can be considered a systematic literature review. Michiels et al. (2016) identified all the publications that fulfil a set of criteria (randomised controlled trials, phase II or III, that recruited HER2+ MBC patients in 1992–2008, and where at least one of the study arms investigated an HER2-targeted agent), and via a collaboration with industrial partners, included those studies for which IPD was available from industry-led studies. This represents an example of possible agreement where the industry could support the availability of IPD.

1.1.3. Definition of PFS and other measures included in the analysis

One of the main issues in the analysis of surrogacy is the heterogeneity related to the definition of progression among clinical trials. This heterogenity can be observed for the period within which patients are evaluated; time intervals between radiologic and clinical

assessments; and the criteria applied to consider patient progression (e.g. variation of the size of the tumour). This heterogenity can be even more important if we consider some tumours' characteristics. For example, as mentioned by Kawakami et al. (2013), in those trials that analysed measurable lesions, the progression is normally documented by radiological assessment (e.g., Response Evaluation Criteria In Solid Tumors, RECIST or World Health Organisation, WHO criteria). In those trials related to non-measurable lesions, disease progression does not necessarily need radiological assessment. Nine out of the 32 articles that use aggregate clinical trial information did not consider or mention the problem of heterogenity on the definition of PFS. Sixteen articles mentioned the problem as a limitation, but did not adjust the methodology in response to the problem. Two studies included only clinical trials that have the same set of progression criteria (RECIST criteria) (Flaherty et al., 2014; Terashima et al., 2015). Three articles considered the problem during the sensitivity analysis where variables such as presence of measurable lessions and tumour response are included (Adunlin et al., 2015; Han et al., 2013; Kawakami et al., 2013). Finally, two authors used established definitions to extract the information collected from the clinical trials, regardless of the terminology used by the original authors (Chen et al., 2015; Petrelli and Barni, 2014).

Moreover, clinicians possibly introduce variance into the measurements of PFS since they decide on the date on which responses were recorded. This could affect studies based on clinical trial information as well as studies based on IPD (Imai et al., 2015; Yoshino et al., 2014)

In addition to the problem of heterogeneity, the literature suggests that there is not always clear information in the clinical trial reports as to how disease progression was evaluated (Hotta et al., 2013; Shitara et al., 2012; Shitara et al., 2014). This and the problem of heterogeneity indicate that there is a need to standardise clinical trial protocols to provide comparability between trials in the same cancer type. Moreover, more complete clinical trial reports that allow access to all information should prove useful.

In addition, 19 out of the 32 of the studies based on trial data combine PFS and TTP into a single surrogate measure. PFS differs from TTP in that, in addition to progression, PFS includes death as a result of any cause while TTP is censored. This means that TTP is the same as PFS only when death does not occur during treatment. All-cause mortality can dilute the association between PFS/TTP and OS. Seven of the 19 articles that combine PFS and TTP analyse the sensitivity of the results by breaking down the articles into those that measure PFS and those that measure TTP. Delea et al. (2012), (Petrelli and Barni, 2014) and (Shitara et al., 2012) found a higher correlation among studies that include PFS in comparison that those that include TTP while (Moriwaki et al., 2016) found a slightly lower correlation when TTP trials were excluded.

Finally, apart from PFS, the literature suggests other key possible surrogates for OS: post-progression survival (PPS), response rate (RR), disease control rate (DCR), time to progression (TTP) and disease-free survival (DFS). 29 out of the 48 selected articles that use clinical trial and/or IPD information analysed more than one surrogate measure. From these 29 articles 16 mentioned PPS.

1.1.4. Factors that affect the relationship between PFS and OS

Based on the variables included as part of the sensitivity analysis or that have been included in the multivariate analysis, we identify a group of factors that the literature

indicates could affect the relationship between PFS and OS. These factors are listed in Table 9.

Table 9. Factors that authors include in the sensitivity analysis or in the multivariate analysis relationship between OS and PFS*

Factor	Number of Studies	Factor	Number of Studies
Treatment line 1st/2nd/3rd	14	Region (Global or regional)	6
Year of the clinical trial	13	Leave one clinical trial out: outliers	5
Type of therapy (e.g. chemotherapy single or in combination)	13	Inclusion of other surrogates apart from PFS	4
Type of treatment	13	Treated with targeted agents	4
Sub-group of patients or tumour type (including patient risk group)	10	Clinical trial phase I, II, III	3
PFS vs TTP	8	Prior treatment or Newly diagnosed vs recurrent	3
Sample size	8	Landmark analysis	3
Crossover	5	Presence of measurable lesion	3

*Only those factors that appear at least in two studies are listed.
Source: Authors' research

As is observed in Table 9, treatment line is the most commonly analysed factor. As we mentioned above, a reason for the importance of the treatment line is the difficulty of separating the PFS effect from the effect of subsequent lines of treatments on OS. This problem is particularly important for first line treatments that are followed by subsequent treatments. For instance, Petrelli and Barni (2014), who analysed 20 first line clinical trials, proposed that the decreases in correlation between PFS HR and OS HR observed in recent years is likely to be due to the influence of post-progression treatments.

The year in which the clinical trial was conducted or published is part of the list of the five most mentioned factors. Here two arguments explain its importance. First, the number of drugs available has grown considerably for a number of cancer types during the past two decades, which means that patients have a larger number of options after progression (Aboshi et al., 2014). Second, the criteria applied to measure progression have been changed. RECIST was published in 2000 which has been recently modified to become the modified RECIST assessment (mRECIST) (Lencioni and Llovet, 2010).

The importance of the type of treatment and therapy is also recognised. The literature suggests that the relationship between PFS and OS can be different within the same cancer trial depending on the treatment applied or the therapy selected.

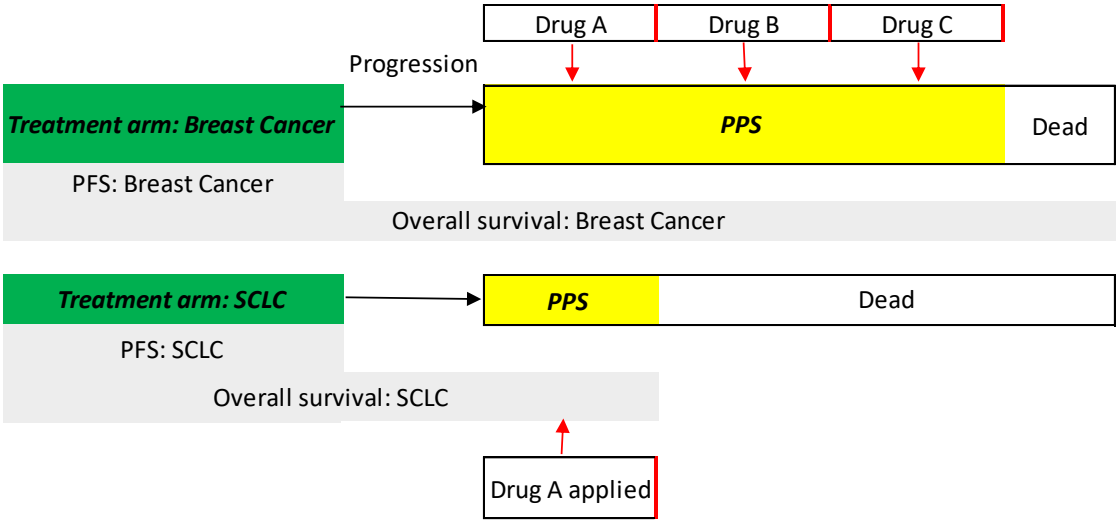
A key factor pointed out by the researchers is the possibility of crossover among trial arms. Five articles that use clinical trial information considered crossover in the sensitivity analysis or in the multivariate analysis. In an analysis of colorectal cancer trials, Adunlin et al. (2015) found that among crossover trials the strength of the association between PFS and OS was higher and more consistent. However, for renal cell carcinoma, the results stemming from Delea et al. (2012) indicate that the link between the effect of the treatment on PFS and the effect of the treatment on OS was stronger in studies that did not allow crossover. In melanoma, Flaherty et al. (2014) suggested that correlation coefficients for the nine trials without crossover were more than 7% higher

than with crossover, and 4% higher when two additional articles that allowed 50% crossover were included. Hotta et al. (2013) examined how crossover therapy affects the relationship between PFS and OS in NSCLC. For clinical trials in which the median proportion of crossover was zero or lower than 1%, the association between the HR of PFS and OS was strong. Kim and Prasad (2016) evaluated previous publications to assess the strength of the surrogate-survival correlation among cancer drugs approved, based on a surrogate endpoint for the USFDA. In Kim and Prasad (2016), there were no significant differences in survival benefit between clinical trials with or without crossover. They suggest that the results are opposed to the commonly-shared idea that crossover masks OS benefits, possibly because crossover prevents observation of late toxicity. Additionally, although Singh et al. (2014) did not include crossover during the sensitivity analysis, they overcame the problem by excluding two clinical trials, where a high proportion of crossoverr were observed, from the quantitative estimation. In addition to the six articles mentioned, 18 out of the 32 articles (50%) that use aggregate clinical trial information mentioned crossover while 8 did not consider or mention crossover. These results provide evidence that researchers recognise crossover as a problem that could affect the validation of PFS as a surrogate measure.

The geographical context in which the clinical trial is conducted (global or regional) is mentioned as a possible factor affecting the validation of PFS as a surrogate measure. A reason to explain this is the variation between Asian and occidental countries in relation to the standard treatments used. This variation implies that clinical trials in both regions used different comparators. In addition, Shitara et al. (2012) pointed out that in the case of advanced gastric cancer, a number of differences in tumour characteristics and practice patterns (e.g. in surgery and chemotherapy) have been identified in Asian and occidental countries.

An important number of the factors presented in Table 9 relate to the characteristics of post-progression survival (treatment line 1st/2nd/3rd, year of the clinical trial, crossover between control and treatment arms, newly diagnosed vs recurrent, and sub-group of patients). These and the fact that an important number of articles analysed PPS together with PFS suggest that PPS has a role in the discussion of the validation of PFS as a surrogate of OS. Figure 2 is an ilustration of how the length of PPS affects the validation of PFS as a surrogate measure. If those patients that suffer from advance SCLC cancer have only a limited number of options for subsequent treatment, we would expect a short PPS. On the other hand, patients that suffer from metatastic breast cancer have several treatment options after progression and a longer OS. Therefore, as ilustrated in Figure 2, the probability that the relationship between the PFS and OS is obscured by the factors related to the PPS is higher for breast cancer.

Figure 2. Effect of PPS on the relationship between PFS and OS



The influence of PPS was confirmed by the study conducted by Amir et al. (2012). Their results indicate that when PPS is short, the correlation between OS and PFS is smaller than when PPS is long. Similarly, for patients with advanced NSCLC, Suzuki et al. (2015) identified the optimal point of correlation of the HR for PFS and the HR for OS by analysing every 1 month of SPP. They found that the correlation between the HR for PFS and for OS decreases for a PPS of less than 6 months. From the 16 articles that analysed PPS, 13 suggest that the relationship between OS and PPS is stronger than between OS and PFS and one of the remaining three pointed out a high correlation among PFS and PPS. These results suggest that subsequent treatments have additional effects on OS than the treatment under consideration.

Consequently, a group of Japanese researchers have become particularly interested in determining the factors that affect the relationship between PPS and OS. Imai et al. (2014), Imai et al. (2015), Kasahara et al. (2015), Yoshino et al. (2014) and (Yoshino et al., 2015) analysed the prognostic factors for PPS using IPD (Table 6). In general, their conclusions suggest that the significant factors to explain the effect of PPS on OS are:

- Number of regimens employed after progression
- Response to the second or third-line treatment (progressive disease (PD)[2] vs non-PD)
- Performance status at progression
- PFS of first line chemotherapy
- Tumour stage after initial treatment
- The presence of distant metastases at recurrence.

These results and future investigations in the topic will be crucial for the analysis of the surrogacy of PFS. To the extent to which protocols of future follow-up of clinical trial patients consider procedures for gathering information of factors that reflect the effect of the post-progression management of the disease, it will be possible to extract stronger

[2] Progressive disease (PD): ≥ 20% increase in the total diameter of all target lesions relative to the smallest total diameter observed during the study.

conclusions from the statistical analysis to validate PFS as a surrogate of OS. Therefore, future research on this topic and its incorporation into clinical trials ishould be strongly encouraged.

Moreover, given that new treatments are regularly being approved, these results also suggest the need for reviewing and updating the earlier studies in the light of the presence of fundamentally-different conditions.

1.1.5. Summary of the literature review

The analysis suggests that it is crucial to increase the use of IPD. As in Davis et al. (2012), our findings suggest that there is still a limitation in the availability of such information. Only one third of the selected articles since 2012 include IPD, and judging from these articles, much of the available IPD information was not extracted from the clinical trials being reported. Using only a single estimate for PFS and a single estimate for OS from a reported trial wastes an enormous amount of data.

The percentage of articles that concluded that PFS is an appropriate surrogate for OS (45%) is higher than the percentage of those that do not support surrogacy (25%). However, our results suggest that the conclusions vary considerably among cancer types. Moreover, an additional 24% of the sample suggests that surrogacy depends on factors such as the length of the PPS and whether the treatment was first or subsequent line. Additionally, there are important variations in the statistical estimation methodology used to support or reject the surrogacy. These variations are observed both within and between cancer types.

Similar to the findings of Davis et al. (2012), correlation (particularly Spearman, Pearson and Kendall's τ) and weighted or unweighted linear regression remain the most common statistical methodologies applied. In addition, a methodology that has increased in importance is the estimation of the STE. This represents the lower bound of PFS that can predict that OS is in the right direction with a sufficient degree of certainty.

However, variations in methodology make the comparison of results difficult. There is a high variation in the characteristics of the methodologies and there is no consistency ln what should be considered appropriate statistical estimation methodology to support the validity of PFS as a surrogate measure for OS. There is an urgent need for standardisation that allows for more consistent results. This would facilitate the use of PFS by policy makers.

The presence of outliers as well as the absence of a linear regression assumption test, lead us to believe that there has been a lack of rigour in a number of applications of linear regression methodology. The importance that validating PFS as a surrogate for OS has on allowing patients to access new health technologies more quickly should not be undermined by a poor knowledge of the methodology applied. It is crucial that authors correctly consider the characteristics of their data, since not doing so could lead to poor or even wrong conclusions. Similarly, policy makers who use this information to decide on whether PFS is an appropiate surrogate for OS should also have the knowledge to understand the weaknesses of poor study design, and apply appropriate methodology rigorously.

In addition to heterogenity in the methodology and interpretation of the results, a further issue in the analysis of surrogacy is the heterogeneity relating to the definition of progression among clinical trials. This and the difficulty in finding the required information in the clinical trial reports indicates a need for standardisation of clinical trial

protocols that allows for comparability between trials in the same cancer type. However, the experts highlighted the improvement in the accuracy of PFS measurement due to the use of PET and MRI scans which, although this does not resolve the problem of heterogenity in the definition, will reduce the measurement errors.

The literature suggests factors that affect the validation of surrogacy. Many of these factors are related to the length and characteristics of post-progression survival. These results and future investigations into the topic will be crucial for the analysis of the surrogacy of PFS. Procedures for gathering information on factors affecting the post-progression management of a disease should be described in protocols for following-up clinical trial patients. If this were done, it should make it possible to derive stronger conclusions from statistical analysis that could validate PFS as a surrogate of OS. Future research on this topic and its incorporation into clinical trials should be strongly encouraged.

It is important to highlight that we have not conducted a systematic literature review of any particular cancer type. Therefore, we cannot say whether PFS should or should not be used in any particular case.

3. WHAT THE EXPERTS SAY – FUTURE OF THE ANALYSIS OF PFS AS A SURROGATE FOR OS

The literature review has told us what tends to work well, what appears not to work, and the existence of a grey area between the two. It says very little about the institutional framework in which decisions are made. Making decisions has a subtly different methodology, accounting for factors that are not amenable to RCTs and which are lumped into an amorphous group that we know as "experience". Often those with experience are not always sure of what the factors are or their relative importance, but both they and non-experts recognise that this knowledge is valuable but cannot easily be measured. These factors will include, formally or informally, the value for money of the interventions, and the ability to pay for them, in deciding those to be recommended.

1.2. Interviews and workshop

We identified a group of leading researchers, individuals from a regulatory authority (EMA) and a reimbursement body (NICE), and chose someone from each group to interview. We asked them to describe what they saw as the advantages and disadvantages of surrogate endpoints, the potential for their improvement and what was needed to take the concept further (i.e. a future research agenda). We conducted interviews with three researchers, one person from EMA and conducted an email correspondence with two people from NICE.

A 3-hour workshop was facilitated by the OHE to expand on the topics identified in the interviews. It was attended by all but one of the experts interviewed, a NICE Appraisals Committee Chair and a number of POI members.

In the following section, we summarise the opinions of the experts in developing a process to determine an appropriate endpoint for the approval of a new cancer drug.

Overlapping points were made by experts in interviews and correspondence, and similar points were made in the workshop. This section presents a summary of the qualitative information collected throughout the interviews and the workshop.

Experts mentioned that the importance of finding an adequate surrogate is related to the need for adjusting for post-progression treatments which could "muddy the waters". Nevertheless, the qualitative analysis also suggests that this might not be a huge problem if the post-progression treatments are not the experimental treatment under investigation and if they form part of a realistic treatment pathway. One of the experts mentioned that if the problem is switching from the control treatment to the experimental treatments, there are methods to deal with this and therefore we shouldn't be thinking about the use of surrogates in isolation from these other adjustment methods.

In this regard, among NICE appraisal committees one view is that PFS may sometimes be a better measure than OS because it is not confounded by post-progression treatments. A separate view expressed by NICE representatives was whether crossovers may nullify effectiveness analysis (and also cost effectiveness). If there are post-progression switches to the experimental treatment under investigation, then the NICE decision problem is confounded. Then we need to use adjustment methods to compare a state of the world in which the new treatment exists with one in which it does not exist. Post-progression switching onto *other* available treatments could reduce the problem,

because such switches may simply reflect what occurs in reality and thus subsequent OS data are valid and do not need to be adjusted.

A further view suggested that we should be still very interested in OS, because it measures the effect of the totality of all the treatments. The NICE decision problem should usually be "what is the cost-effectiveness of inserting this treatment into the treatment pathway" rather than "what is the cost-effectiveness of this treatment if we make the unrealistic assumption that patients receive no post-progression treatments".

In addition, concerns were raised regarding the lack of follow-up that could hinder the extrapolation of the information. Moreover, there was also a concern about the impossibility of correctly measuring OS since, as one expert commented, "there is no OS if only few deaths have occurred".

As a result, although follow-up may not be complete, the data are not totally unobserved and therefore we can use the data we have to conduct extrapolation. If extrapolation is not considered appropriate, that may be due to extrapolation techniques not being good enough. In addition, the view was expressed that modelling for PFS and for OS may sometimes be carried out in a way that best suits the modeller.

Other aspects mentioned by the experts that could hinder extrapolation was the accuracy of the PFS measurement. Nevertheless, one expert pointed out that with the advent of more accurate measurements, in particular due to the wider use of PET and MRI scans, the accuracy of a number of surrogate endpoints has improved, which reduces the measurement errors for the surrogate endpoints for solid cancers. This reduction in noise translates into higher correlation between the surrogate and the final endpoint.

Even when the techniques used to extrapolate PFS to OS are appropriate, an important criticism of the use of PFS as a surrogate measure is that it does not reflect the QoL of the patient. However, the opinion of some experts is that PFS may reflect the endpoint "quality of life" just as well as DFS. In a number of cases, PFS can be regarded as one of the "true" or final endpoints, to the extent that it may often be a good proxy for quality of life. For example, in lung cancer, an increase in tumour size from 2 cm to 2.4 cm (the distinction between DFS and PFS) will not usually feel different to the patient, and the difference will not usually be clinically relevant.

To the extent that all measures of QoL are subjective, there is no gold standard for QoL, so the correlation between PFS and QoL need not be high. In addition, and probably more importantly, it is more likely that PFS and the component of a QALY that depends on QoL should be better correlated, because the latter will depend on the length of time that the treatment is effective (as well as the QoL during that time). The length of time that the treatment is effective could be measured a number of ways, including by PFS itself.

In cancer patients with a low life expectancy, and where treatment is unlikely to make much, if any, difference in the length of remaining life, an improvement in the quality of remaining life may be more important. In this case, PFS and OS may not be highly correlated (because QoL as measured by PFS has changed a lot and will probably not be highly correlated with small changes in OS). Unlike OS, which is a gold standard for increased life expectancy, quality of life has no obvious gold standard, because quality of life is measured subjectively, so agreement on how to measure it may be carried out to achieve consistency of measurement rather than reflect an objective measure.

Another concern expressed by the experts was whether surrogate endpoints are treated too leniently by regulators and reimbursement bodies. A comparison of a large number of trials in which surrogate endpoints have been used has shown that the treatment is successful relatively more frequently than for trials in which a surrogate endpoint was not employed. Other things being equal, it would imply that surrogate endpoints are relatively optimistic. If OS is the gold standard, then the use of a surrogate would, if anything, suggest that because of the additional uncertainty about the final endpoint by its prediction by the surrogate, the trials using a surrogate should demonstrate a lower level of success than those that do not.

- A number of recent treatment advances have not followed the same pattern as many previous treatments, the object of which has been to delay disease progression for a time, but then become ineffective. New treatments include the introduction of an "imperfect surrogate". In breast cancer, the introduction of chemotherapy before surgery greatly improves the chances of survival – and indeed, of cure. For women aged around 40 years of age, the cure rate rises to about 70%. In such circumstances, it is of little account that the extent of the shrinkage due to chemotherapy might not be well correlated with overall survival.
- Minimal residual disease in pathologies outside oncology. With human immunodeficiency virus (HIV), the surrogate "viral load" may be undetectable, which does not mean that the patient has been cured, but it does mean that taking Highly Active Antiretroviral Therapy (HAART) for life will result in an almost-normal lifespan. The correlation of something that is too small to be measured (viral load) with OS is unknown, but given the outcome, that is not material. Something similar occurs now with some leukaemia treatments. Again, the form of validation of the surrogate as a predictor of OS will need to change to reflect the reality of success.
- The use of liquid biopsies has brought a possibility for a new surrogate. Cell-free tumour DNA (ctDNA) or fragments of it break off the tumour and circulate in the bloodstream. Some of these pieces of ctDNA become useful as biomarkers for different forms of cancer, from which predictions can be made about the stage and spread of the cancer, as well as allowing a capability for monitoring the effect of treatment. That is, given that it can both predict the effect of treatment and also the probability of overall survival, ctDNA is in effect an agent that influences surrogate endpoints and also OS. As it can be gained non-invasively from a blood test, it can take the place of more invasive forms of biopsy.
- In immunotherapy, the body produces its own defences against cancer. Immunotherapy assists the body to improve its defences in some way. A number of such therapies have been discovered. The use of standard measures of progression as a surrogate may be particularly inappropriate for these therapies, because "pseudo-progression" has been observed (Pilotto et al., 2015). Therapies may demonstrate very little PFS benefit, but a substantial OS gain. New definitions of "progression" may be required, in place of the commonly used RECIST criteria (Eisenhauer et al., 2009).

Despite these advances, it is also necessary to continue to improve the availability of information in order to more accurately predict OS by using surrogate endpoints to predict OS. Experts agreed with the principle of making individual level data available more freely, provided that the data can be kept anonymous and as long as they are shared responsibly, so that any conclusions follow from established scientific principles.

They noted that uninformed and pseudo analysis would probably do more harm than good. The issue has been the subject of lengthy debates and discussions between governments, drug regulators and payers, academics and the pharmaceutical industry for a number of years. A summary of these arguments is given in (ABPI, 2013). It is beyond the scope of this report to rehearse these arguments further, other than to emphasise that substantial benefits to the health of the population and for the more rapid adoption of healthcare technology could be forthcoming if IPD were to be used in conjunction with surrogate endpoints.

1.3. Technical studies

Two recent technical studies are of relevance to this project. The first, by Buyse et al. (2016) was discussed in the Workshop and at a subsequent POI member meeting. The other study – Stevens et al. (2014) – was not discussed at those meetings, as it was discovered too late. Note that this review has not been systematic in the same way that the literature review of the trial evidence attempted to be.

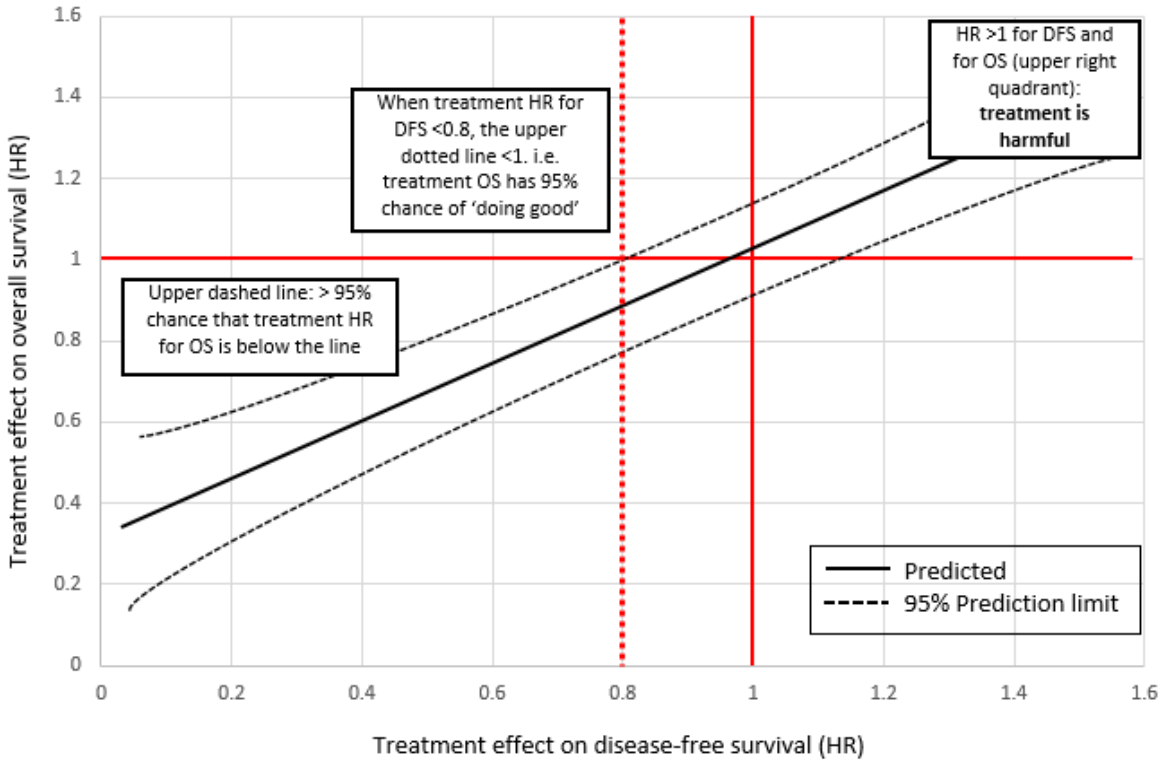
1.4. Buyse et al. (2016)

Buyse et al. (2016) have developed a statistical methodology to determine what constitutes a good surrogate. They describe the history for this methodology, beginning with Prentice (1989). The original set of characteristics for what makes a good surrogate has been found to be incomplete, but each successive attempt to improve the characteristic set has also been shown to be lacking in some circumstances. That has meant progress has been slow, and that trade-offs between different ways of characterising the value of a surrogate need to be made. Buyse et al. (2016) note that, for a single trial, the usefulness of a surrogate depends not only upon the correlation between the surrogate endpoint and the true endpoint, but also between the intervention and the surrogate effect, the intervention and the true effect and an interaction between the intervention and the surrogate effect in predicting the true effect. In such trials, individual-level data are not generally reported or made available, and only the mean effects of the surrogate and the true effect are used.

A further layer of complexity arises when there is more than one trial. Differences between trial designs such as in different dosages of a drug or in the population being treated mean that an individual-trial-level effect and a between-trial-level effect (called an "individual effect" and a "trial effect" respectively by Buyse et al. (2016)) need to be distinguished.

Buyse et al. (2016) define a surrogate threshold effect (STE) as the smallest treatment effect on the surrogate that predicts a nonzero treatment effect on the true endpoint (Figure 3).

Figure 3. Example of the surrogate threshold effect (STE)



Source: Adapted from Buyse et al. (2016)

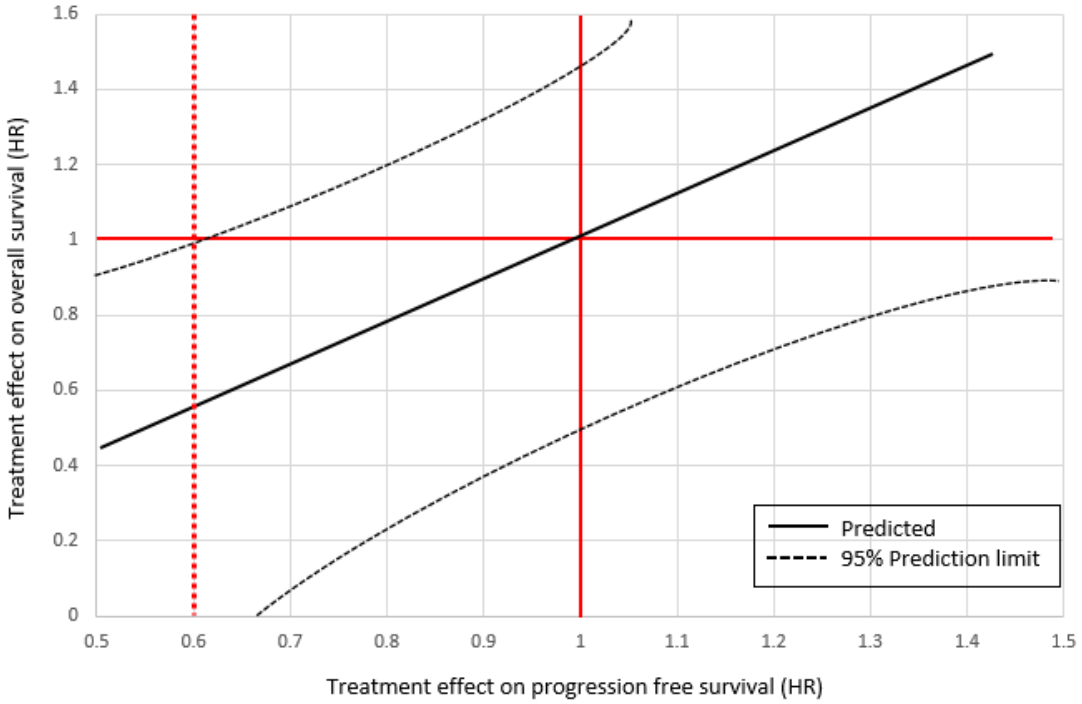
In Figure 3 the dashed line above the line of best fit indicates the upper 95% confidence interval of predicting OS from DFS. DFS with a HR <1 shows that DFS predicts an improvement in overall survival (OS). The vertical axis shows the HR for OS, where below 1 is an improvement. There is a relatively high correlation between DFS and OS. We are interested in finding where the upper 95% CI (the dashed line) meets the horizontal line at an OS HR of 1, because below and to the left of this point, the DFS will indicate that the OS is improved, with a confidence level of at least 95%. The two lines meet where DFS is 0.8, and this is thus the STE. That means that if the DFS for the treatment intervention shows an 20% improvement over its comparator, then the HR for the corresponding OS will be at most 1 (with a CI of 95% or higher). This therefore indicates that the OS for the intervention has increased compared with its comparator.

Thus if the DFS is below 0.8, we can be confident (i.e. at least 95% confident) that OS will have increased due to treatment.

Figure 3 shows a situation similar to the one shown for resectable gastric cancer, which means that the cancer can be operated on, and in many or most cases will not be known to have metastasised. In that case, DFS will often if not usually be for the rest of life, so we could readily expect that DFS would equal OS, which would lead to high correlation between the two variables.

In Figure 4, we look at the relationship between DFS and OS in the case where the prediction interval is much wider. A similar wider prediction interval was found by Buyse et al. (2016) in the case of advanced gastric cancer.

Figure 4. Example of the surrogate threshold effect (STE) with a wider 95% prediction interval



Source: Adapted from Buyse et al. (2016)

Here, the relationship between PFS and OS is not so strong. (Note also that Figure 4 is of PFS whereas Figure 3 is of DFS.) The upper dashed line meets the OS line with a HR of 1 where the HR for DFS is 0.6 (i.e. the STE). That implies that the HR for DFS has to be reduced by quite a large amount (~40%) before it can be said with sufficient confidence that the treatment will improve OS.

During the discussion in the workshop, with experts interviewed and in a subsequent meeting with some of the POI members, the following points were made:

- The method of Buyse et al. (2016) requires a number of trials to generate enough data. It is only useful for about 5 or 6 of the cancers that have been studied sufficiently.
- The loss of data due to not using individual level data results in a fall in precision, often leading to the rejection of the surrogate by regulators and reimbursement bodies, because of the uncertainty due to too few data points being available.
- Use of individual data could lead to Buyse et al. (2016)'s method being used for a wider range of cancers.

Two opinions (one associated with NICE) suspected that Buyse's criterion for a good surrogate endpoint may be too stringent. There are other sources of evidence, which, if in the right direction, would allow a lower confidence interval than the 95% CI that Buyse employs. One is that the CI almost certainly does not require a lower bound, and the CI obtained is thus at 97.5%. Another is that we would know, mostly from past experience, how often approved treatments in oncology have reduced overall survival. We would also have some understanding from trial and other uses of relatively new drugs what the adverse event profile of an oncology treatment is likely to be. And we

may need to have a higher regard for the cost effectiveness elements rather than the effectiveness elements for the adoption of the technology.

According to the experts interviewed, the statistical work of Buyse et al. (2016) has had relatively little application so far, despite being available in earlier forms for over a decade. Greater recognition of appropriate statistical methods is required. Ordinary least squares (OLS) and linear regression methods cannot tell the full story and meta-regression may be more promising. This suggests that the attitudes of reimbursement bodies such as NICE in England, IQWIG in Germany and PBAC in Australia towards surrogacy might be more important to analyse and improve than increasing the application of existing statistical methods.

Nevertheless, role of the STE in the analysis of PFS as a surrogate of OS has been increasing. As mentioned by Ciani et al. (2016), to enable prediction of the effect of the treatment on OS based on the value of PFS is key for decisions on coverage and reimbursement since reimbursement bodies must take into account the benefits and costs of and intervention compared with other treatments. In this regard, the opinion of NICE representatives was that the threshold proposed by Buyse et al. (2016) could be useful for regulators. However, for modelling purposes, a point estimate of OS HR would normally be needed.

1.5. Stevens et al. (2014)

The object of the paper is to attempt to find out the benefit to society of being able to bring a product onto the market earlier because the use of surrogate endpoints reduces the time to conduct the trials of cancer drugs and other treatments, and to weigh up those benefits against the cost of using an inaccurate measure of benefit (that is, an inaccurate measure of overall survival). Occasionally, using a surrogate endpoint, a drug or other treatment that on average shortens overall survival compared with usual treatment will be licensed and be put on a reimbursement schedule. Additionally, because a surrogate endpoint has been used, a treatment will sometimes be licensed and put on reimbursement schedules even though it does not do as much good per life-year or QALY gained as a different treatment that could have had a greater effect but which had not been accepted. These things will still happen when the appropriate final endpoint has been used, but can be expected to happen less often.

The study undertook an analysis for the USA and for the EU, and estimates that using current practices, surrogate endpoint use would lead to a net gain in health that has been subsequently been turned into a money gain at a societal rate of $US150,000 per QALY and a European rate of €85,000 per QALY.

The figures using a societal value of a QALY differ from a government willingness-to-pay by a factor of about 3 or 4 to 1(UK Department of Health, paragraph 5.24, 2010), so in conventional terms of the willingness to pay by a national health service, the costs and benefits described above should both be reduced by a factor of 3 or 4.

The study is unusual in that the delay in bringing a treatment onto the market is not normally included in a cost effectiveness or cost benefit study of healthcare. However, for the case of surrogate endpoints, it may be argued that it is the comparison that is most appropriate. An alternative way of including the benefit derived from earlier adoption (and which is recognised in the health economic literature) may be derived using value of information (VOI) analysis (Griffin et al., 2011; Palmer and Smith, 2000)

This study considered the case of metastatic renal cell carcinoma. The authors estimate that the use of a surrogate endpoint (PFS) gave rise to a net saving of life-years, but it is not obvious whether this result would be repeated in other cancers (though others could now conceivably use similar methodology to estimate the relationship between PFS and OS for other oncology areas). The size of the health gains (as estimated) is impressively large, which suggests that the use of surrogates and a two-stage or managed entry type procedure for the reimbursement of new treatments is worthy of further consideration. If the results are repeated in a similar fashion for other cancers, or other stages of a particular cancer, and for different drugs, it could alter the balance in favour of surrogacy in many areas.

Further analysis of the Stevens et al. paper is warranted. The analytical methods used by the authors are open to question and may be flawed, which may invalidate their findings. In addition, the authors have assumed away the problem of accounting for the effects of therapies after the first line therapy by simply saying that this was not considered. To that extent, this is an incomplete study. It requires further work to sort out that problem, if indeed it can.

4. CONCLUSIONS

In the field of surrogate endpoints, what must first be guarded against is that a treatment does not reduce OS except where the express use of a treatment is palliative (that is, to increase the quality of life of patients in their last days) or some similar use of a drug to improve quality of life in the proximity of death. If a surrogate endpoint shows an increase in PFS, the main way that we can discover if OS is also positive is to follow each patient up during their full lifetime, and be assiduous about finding adverse events.

However, treatment should not normally cause a decline in OS (although care needs to be taken to ensure that treatments which provide only small increases in OS are not offset by the effect of treatment toxicity on QoL). The problem will more usually be that a treatment does not increase OS by enough to be significant, that the increase in OS cannot be attributed to a given cause, or that the true value of an increase in OS is too small to be cost effective. Furthermore, the analyses of cost effectiveness with the exception of Stevens (2014) appear to omit the benefits of earlier adoption of technology by employing a surrogate endpoint, or fail to quantify them. Where IPD is available, it appears that value of information analysis may be the best way of determining whether early adoption of a drug (by using a surrogate endpoint) is optimal. When this form of data is not available, consideration of Stevens' approach may prove useful.

Cancer treatment shows that in two respects, standard methodology appears to require some adaptation, in that it does **not**:

- Recognise the benefits of earlier availability of new cancer drugs using surrogate endpoints as a measure of treatment benefit; and
- Cope with the many lines of treatment.

Current systems for deciding upon the allocation of health care resources primarily use a "one-time" procedure. For instance, the novel therapy is assessed based upon the data available (using whatever extrapolation techniques are deemed suitable) and a decision is made as to whether or not it should be provided by the health system. If the benefits of earlier decision making were to be incorporated into an analysis, the system itself

would need to allow "early" and "late" decisions, whether it be through a phased process or a managed entry process. Then, at each stage, a decision maker could opt to grant early access or no access pending a follow-up appraisal at a specified time-point when more information would be available. This has some similarities to the processes outlined for the Cancer Drugs Fund (Mayor, 2016).

The failures of current approaches to the problems raised by surrogacy would suggest several lines of research. One option is an editorial or opinion piece in a top-level generalist or oncology journal, written jointly by clinicians, researchers and economists, pointing out the deficiencies of current evaluation methods. Another could be to extend the work begun by Stevens et al (2014) to other cancers, and to look at exploring different patient populations (by stage of cancer, perhaps) and different drug regimens. Further research work would need to recognise the limits of current methodology and to explore the possibilities of the use of decision theory for both the effectiveness and the cost effectiveness stages of an analysis (at present, decision theory is only used for the cost effectiveness stage).

One line of research that could address the problem recognised above of "OS does not increase by enough to be significant" would be to carry out the analysis of data using IPD. The problems would seem to be to ensure that anonymity of IPD data could be maintained, that IPD data could only be accessed by bona fide researchers and that manufacturers providing such data should not be put at a disadvantage compared with manufacturers who do not offer data. Data sources such as clinicalstudydatarequest.com represent a good start to the more widespread sharing of data, but further advances are necessary. Overall, improving the accuracy of surrogacy should reduce the lead time before a drug can be licensed and therefore be of use to both patients and manufacturers.

Some of these changes, if implemented, could significantly reduce the delays in the adoption of cost-effective new technology. If so, this should improve both population health and manufacturer profits as well as increase incentives for innovation in health technologies.

The literature review shows that the recognition of the issues described above is slowly improving, but that the speed of change has been hindered by the lack of application of a common code of reporting practice. The spread of the use of STE demonstrates that there has been some recognition of issues wider than the correlation of PFS (or other surrogate endpoint) and OS. The following limitations of the literature review should be acknowledged. First, our analysis was based on a citation search, which in turn was based on the 19 articles found by Davis et al. (2012) and is not a systematic literature review. This means that we are assuming that the search done by Davis et al. (2012) was able of identify the main literature published until 2012 on the evidence of PFS as an appropriate surrogate for OS. In addition, we did not include any grey literature. Nevertheless, as mentioned by Davis et al. (2012), a systematic literature review of the topic would be infeasible.

In summary, the use of surrogate endpoints for OS in oncology has the potential for facilitating early access of patient to new compounds. However, it is necessary to understand which factors affect the relationship between PFS and OS. In this regard, more research on the determinants of PPS is required as well as an improvement in the accessibility to IPD information collected during the RCT. This requires strong collaboration between the industry, payers and regulators to establish standards that

support an increase in the quality of the research and ensure that the evolution of the main factors that affect the relationship is considered. Researchers also need to be aware of technical advances being attempted in statistical and economic methodologies in this area.

5. REFERENCES

5.1 References – Literature review

Aboshi, M., Kaneko, M. and Narukawa, M., 2014. Factors affecting the association between overall survival and progression-free survival in clinical trials of first-line treatment for patients with advanced non-small cell lung cancer. Journal of Cancer Research and Clinical Oncology, 140 (5), 839-848.

Adunlin, G., Cyrus, J. W. W. and Dranitsaris, G., 2015. Correlation between progression-free survival and overall survival in metastatic breast cancer patients receiving anthracyclines, taxanes, or targeted therapies: a trial-level meta-analysis. Breast Cancer Research and Treatment, 154 (3), 591-608.

Agarwal, N., Bellmunt, J., Maughan, B. L., Boucher, K. M., Choueiri, T. K., Qu, A. Q., Vogelzang, N. J., Fougeray, R., Niegisch, G., Albers, P., Wong, Y.-N., Ko, Y.-J., Sridhar, S. S., Tantravahi, S. K., Galsky, M. D., Petrylak, D. P., Vaishampayan, U. N., Mehta, A. N., Beer, T. M., Sternberg, C. N., Rosenberg, J. E. and Sonpavde, G., 2014. Six-Month Progression-Free Survival as the Primary Endpoint to Evaluate the Activity of New Agents as Second-line Therapy for Advanced Urothelial Carcinoma. Clinical Genitourinary Cancer, 12 (2), 130-137.

Amir, E., Seruga, B., Kwong, R., Tannock, I. F. and Ocaña, A., 2012. Poor correlation between progression-free and overall survival in modern clinical trials: Are composite endpoints the answer? European Journal of Cancer, 48 (3), 385-388.

Beauchemin, C., Cooper, D., Lapierre, M.-È., Yelle, L. and Lachaine, J., 2014. Progression-free survival as a potential surrogate for overall survival in metastatic breast cancer. OncoTargets and therapy, 7 1101-1110.

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

Bowater, R. J., Lilford, P. E. and Lilford, R. J., 2011. Estimating changes in overall survival using progression-free survival in metastatic breast and colorectal cancer. International journal of technology assessment in health care, 27 (03), 207-214.

Bria, E., Massari, F., Maines, F., Pilotto, S., Bonomi, M., Porta, C., Bracarda, S., Heng, D., Santini, D., Sperduti, I., Giannarelli, D., Cognetti, F., Tortora, G. and Milella, M., 2015. Progression-free survival as primary endpoint in randomized clinical trials of targeted agents for advanced renal cell carcinoma. Correlation with overall survival, benchmarking and power analysis. Critical Reviews in Oncology/Hematology, 93 (1), 50-59.

Burzykowski, T., Buyse, M., Piccart-Gebhart, M. J., Sledge, G., Carmichael, J., Lück, H.-J., Mackey, J. R., Nabholtz, J.-M., Paridaens, R. and Biganzoli, L., 2008. 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, 26 (12), 1987-1992.

Buyse, M., Burzykowski, T., Carroll, K., Michiels, S., Sargent, D. J., Miller, L. L., Elfring, G. L., Pignon, J.-P. and Piedbois, P., 2007. Progression-free survival is a surrogate for survival in advanced colorectal cancer. Journal of Clinical Oncology, 25 (33), 5218-5224.

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

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., der Elst, W. and Burzykowski, T., 2016. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. Biometrical Journal, 58 (1), 104-132.

Cartier, S., Zhang, B., Rosen, V. M., Zarotsky, V., Bartlett, J. B., Mukhopadhyay, P., Wagner, S. and Davis, C., 2015. Relationship between Treatment Effects on Progression-Free Survival and Overall Survival in Multiple Myeloma: A Systematic Review and Meta-Analysis of Published Clinical Trial Data. Oncology Research and Treatment, 38 (3), 88-94.

Chen, Y.-P., Sun, Y., Chen, L., Mao, Y.-P., Tang, L.-L., Li, W.-F., Liu, X., Zhang, W.-N., Zhou, G.-Q., Guo, R., Lin, A.-H. and Ma, J., 2015. Surrogate endpoints for overall survival in combined chemotherapy and radiotherapy trials in nasopharyngeal carcinoma: Meta-analysis of randomised controlled trials. Radiotherapy and Oncology, 116 (2), 157-166.

Chirila, C., Odom, D., Devercelli, G., Khan, S., Sherif, B. N., Kaye, J. A., Molnár, I. and Sherrill, B., 2012. Meta-analysis of the association between progression-free survival and overall survival in metastatic colorectal cancer. International Journal of Colorectal Disease, 27 (5), 623-634.

Ciani, O., Buyse, M., Garside, R., Peters, J., Saad, E. D., Stein, K. and Taylor, R. S., 2015. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. Journal of Clinical Epidemiology, 68 (7), 833-842.

Ciani, O., Davis, S., Tappenden, P., Garside, R., Stein, K., Cantrell, A., Saad, E. D., Buyse, M. and Taylor, R. S., 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), 312-324.

Davis, S., Tappenden, P. and Cantrell, A., 2012. A review of studies examining the relationship between progression-free survival and overall survival in advanced or metastatic cancer. Sheffield: Decision Support Unit, ScHARR, University of Sheffield.

Delea, T. E., Khuu, A., Heng, D. Y., Haas, T. and Soulieres, D., 2012. Association between treatment effects on disease progression end points and overall survival in clinical studies of patients with metastatic renal cell carcinoma. Br J Cancer, 107 (7), 1059-1068.

Félix, J., Aragão, F., Almeida, J. M., Calado, F. J., Ferreira, D., Parreira, A. B., Rodrigues, R. and Rijo, J. F., 2013. Time-dependent endpoints as predictors of overall survival in multiple myeloma. BMC cancer, 13 (1), 1.

Flaherty, K. T., Hennig, M., Lee, S. J., Ascierto, P. A., Dummer, R., Eggermont, A. M. M., Hauschild, A., Kefford, R., Kirkwood, J. M., Long, G. V., Lorigan, P., Mackensen, A., McArthur, G., O'Day, S., Patel, P. M., Robert, C. and Schadendorf, D., 2014. Surrogate endpoints for overall survival in metastatic melanoma: a meta-analysis of randomised controlled trials. The Lancet Oncology, 15 (3), 297-304.

Fleischer, F., Gaschler-Markefski, B. and Bluhmki, E., 2009. A statistical model for the dependence between progression-free survival and overall survival. Statistics in Medicine, 28 (21), 2669-2686.

Foster, N. R., Qi, Y., Shi, Q., Krook, J. E., Kugler, J. W., Jett, J. R., Molina, J. R., Schild, S. E., Adjei, A. A. and Mandrekar, S. J., 2011. Tumor response and progression-free survival as potential surrogate endpoints for overall survival in extensive stage small-cell lung cancer. Cancer, 117 (6), 1262-1271.

Foster, N. R., Renfro, L. A., Schild, S. E., Redman, M. W., Wang, X. F., Dahlberg, S. E., Ding, K., Bradbury, P. A., Ramalingam, S. S., Gandara, D. R., Shibata, T., Saijo, N., Vokes, E. E., Adjei, A. A. and Mandrekar, S. J., 2015. Multitrial Evaluation of Progression-Free Survival as a Surrogate End Point for Overall Survival in First-Line Extensive-Stage Small-Cell Lung Cancer. Journal of Thoracic Oncology, 10 (7), 1099-1106.

Galsky, M. D., Krege, S., Lin, C. C., Hahn, N., Ecke, T., Moshier, E., Sonpavde, G., Godbold, J., Oh, W. K. and Bamias, A., 2013. Relationship between 6-and 9-month progression-free survival and overall survival in patients with metastatic urothelial cancer treated with first-line cisplatin-based chemotherapy. Cancer, 119 (16), 3020-3026.

Garon, E. B., 2012. Issues surrounding clinical trial endpoints in solid malignancies with a focus on metastatic non-small cell lung cancer. Lung Cancer, 77 (3), 475-481.

Giessen, C., Laubender, R. P., Ankerst, D. P., Stintzing, S., Modest, D. P., Mansmann, U. and Heinemann, V., 2013. Progression-free survival as a surrogate endpoint for median overall survival in metastatic colorectal cancer: literature-based analysis from 50 randomized first-line trials. Clinical Cancer Research, 19 (1), 225-235.

Giessen, C., Laubender, R. P., Ankerst, D. P., Stintzing, S., Modest, D. P., Schulz, C., Mansmann, U. and Heinemann, V., 2015. Surrogate endpoints in second-line treatment for mCRC: A systematic literature-based analysis from 23 randomised trials. Acta Oncologica, 54 (2), 187-193.

Hackshaw, A., Knight, A., Barrett-Lee, P. and Leonard, R., 2005. Surrogate markers and survival in women receiving first-line combination anthracycline chemotherapy for advanced breast cancer. British journal of cancer, 93 (11), 1215-1221.

Halabi, S., Rini, B., Escudier, B., Stadler, W. M. and Small, E. J., 2014. Progression-free survival as a surrogate endpoint of overall survival in patients with metastatic renal cell carcinoma. Cancer, 120 (1), 52-60.

Halabi, S., Vogelzang, N. J., Ou, S.-S., Owzar, K., Archer, L. and Small, E. J., 2009. Progression-free survival as a predictor of overall survival in men with castrate-resistant prostate cancer. Journal of clinical oncology, 27 (17), 2766-2771.

Han, K., Ren, M., Wick, W., Abrey, L., Das, A., Jin, J. and Reardon, D. A., 2013. Progression-free survival as a surrogate endpoint for overall survival in glioblastoma: a literature-based meta-analysis from 91 trials. Neuro-Oncology.

Heng, D. Y., Xie, W., Bjarnason, G. A., Vaishampayan, U., Tan, M. H., Knox, J., Donskov, F., Wood, L., Kollmannsberger, C. and Rini, B. I., 2011. Progression-free survival as a predictor of overall survival in metastatic renal cell carcinoma treated with contemporary targeted therapy. Cancer, 117 (12), 2637-2642.

Hotta, K., Fujiwara, Y., Matsuo, K., Kiura, K., Takigawa, N., Tabata, M. and Tanimoto, M., 2009. Time to progression as a surrogate marker for overall survival in patients with advanced non-small cell lung cancer. Journal of Thoracic Oncology, 4 (3), 311-317.

Hotta, K., Kato, Y., Leighl, N., Takigawa, N., Gaafar, R. M., Kayatani, H., Hirata, T., Ohashi, K., Kubo, T. and Tabata, M., 2015. Magnitude of the benefit of progression-free survival as a potential surrogate marker in phase 3 trials assessing targeted agents in molecularly selected patients with advanced non-small cell lung cancer: systematic review. PloS one, 10 (3), e0121211.

Hotta, K., Kiura, K., Fujiwara, Y., Takigawa, N., Hisamoto, A., Ichihara, E., Tabata, M. and Tanimoto, M., 2011. Role of survival post-progression in phase III trials of systemic chemotherapy in advanced non-small-cell lung cancer: a systematic review. PloS one, 6 (11), e26646.

Hotta, K., Suzuki, E., Di Maio, M., Chiodini, P., Fujiwara, Y., Takigawa, N., Ichihara, E., Reck, M., Manegold, C., Pilz, L., Hisamoto-Sato, A., Tabata, M., Tanimoto, M., Shepherd, F. A. and Kiura, K., 2013. Progression-free survival and overall survival in phase III trials of molecular-targeted agents in advanced non-small-cell lung cancer. Lung Cancer, 79 (1), 20-26.

Imai, H., Mori, K., Ono, A., Akamatsu, H., Taira, T., Kenmotsu, H., Naito, T., Kaira, K., Murakami, H., Endo, M., Nakajima, T. and Takahashi, T., 2014. Individual-level data on the relationships of progression-free survival and post-progression survival with overall survival in patients with advanced non-squamous non-small cell lung cancer patients who received second-line chemotherapy. Medical Oncology, 31 (8), 1-7.

Imai, H., Mori, K., Wakuda, K., Ono, A., Akamatsu, H., Shukuya, T., Taira, T., Kenmotsu, H., Naito, T., Kaira, K., Murakami, H., Endo, M., Nakajima, T., Yamamoto, N. and Takahashi, T., 2015. Progression-free survival, post-progression survival, and tumor response as surrogate markers for overall survival in patients with extensive small cell lung cancer. Annals of Thoracic Medicine, 10 (1), 61-66.

Johnson, K. R., Liauw, W. and Lassere, M. N. D., 2015. Evaluating surrogacy metrics and investigating approval decisions of progression-free survival (PFS) in metastatic renal cell cancer: a systematic review. Annals of Oncology, 26 (3), 485-496.

Johnson, K. R., Ringland, C., Stokes, B. J., Anthony, D. M., Freemantle, N., Irs, A., Hill, S. R. and Ward, R. L., 2006. 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, 7 (9), 741-746.

Kasahara, N., Imai, H., Kaira, K., Mori, K., Wakuda, K., Ono, A., Taira, T., Kenmotsu, H., Harada, H. and Naito, T., 2015. Clinical impact of post-progression survival on overall survival in patients with limited-stage disease small cell lung cancer after first-line chemoradiotherapy. Radiology and oncology, 49 (4), 409-415.

Kawakami, H., Okamoto, I., Hayashi, H., Taguri, M., Morita, S. and Nakagawa, K., 2013. Postprogression survival for first-line chemotherapy in patients with advanced gastric cancer. European Journal of Cancer, 49 (14), 3003-3009.

Kim, C. and Prasad, V., Strength of validation for surrogate end points used in the US Food and Drug Administration's approval of oncology drugs. Mayo Clinic Proceedings, 2016. Elsevier, 713-725.

Laporte, S., Squifflet, P., Baroux, N., Fossella, F., Georgoulias, V., Pujol, J.-L., Douillard, J.-Y., Kudoh, S., Pignon, J.-P., Quinaux, E. and Buyse, M., 2013. Prediction of survival benefits from progression-free survival benefits in advanced non-small-cell lung cancer: evidence from a meta-analysis of 2334 patients from 5 randomised trials. BMJ Open, 3 (3).

Lencioni, R. and Llovet, J. M., Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. Seminars in liver disease, 2010. © Thieme Medical Publishers, 052-060.

Li, X., Liu, S., Gu, H. and Wang, D., 2012. Surrogate end points for survival in the target treatment of advanced non-small-cell lung cancer with gefitinib or erlotinib. Journal of Cancer Research and Clinical Oncology, 138 (11), 1963-1969.

Louvet, C., de Gramont, A., Tournigand, C., Artru, P., Maindrault-Goebel, F. and Krulik, M., 2001. Correlation between progression-free survival and response rate in patients with metastatic colorectal carcinoma. Cancer, 91 (11), 2033-2038.

Mandrekar, S. J., Qi, Y., Hillman, S. L., Ziegler, K. L. A., Reuter, N. F., Rowland, K. M., Kuross, S. A., Marks, R. S., Schild, S. E. and Adjei, A. A., 2010. Endpoints in phase II trials for advanced non-small cell lung cancer. Journal of Thoracic Oncology, 5 (1), 3-9.

Matulonis, U. A., Oza, A. M., Ho, T. W. and Ledermann, J. A., 2015. Intermediate clinical endpoints: A bridge between progression-free survival and overall survival in ovarian cancer trials. Cancer, 121 (11), 1737-1746.

Mauguen, A., Pignon, J.-P., Burdett, S., Domerg, C., Fisher, D., Paulus, R., Mandrekar, S. J., Belani, C. P., Shepherd, F. A., Eisen, T., Pang, H., Collette, L., Sause, W. T., Dahlberg, S. E., Crawford, J., O'Brien, M., Schild, S. E., Parmar, M., Tierney, J. F., Pechoux, C. L. and Michiels, S., 2013. Surrogate endpoints for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer: a re-analysis of meta-analyses of individual patients' data. The Lancet Oncology, 14 (7), 619-626.

Michiels, S., Pugliano, L., Marguet, S., Grun, D., Barinoff, J., Cameron, D., Cobleigh, M., Di Leo, A., Johnston, S., Gasparini, G., Kaufman, B., Marty, M., Nekjudova, V., Paluch-Shimon, S., Penault-Llorca, F., Slamon, D., Vogel, C., von Minckwitz, G., Buyse, M. and Piccart, M., 2016. Progression-free survival as surrogate endpoint for overall survival in clinical trials of HER2-targeted agents in HER2-positive metastatic breast cancer. Annals of Oncology.

Miksad, R. A., Zietemann, V., Gothe, R., Schwarzer, R., Conrads-Frank, A., Schnell-Inderst, P., Stollenwerk, B. and Siebert, U., 2008. Progression-free survival as a surrogate endpoint in advanced breast cancer. International journal of technology assessment in health care, 24 (04), 371-383.

Moriwaki, T., Yamamoto, Y., Gosho, M., Kobayashi, M., Sugaya, A., Yamada, T., Endo, S. and Hyodo, I., 2016. Correlations of survival with progression-free survival, response rate, and disease control rate in advanced biliary tract cancer: a meta-analysis of randomised trials of first-line chemotherapy. Br J Cancer, 114 (8), 881-888.

Négrier, S., Bushmakin, A. G., Cappelleri, J. C., Korytowsky, B., Sandin, R., Charbonneau, C., Michaelson, M. D., Figlin, R. A. and Motzer, R. J., 2014. Assessment of

progression-free survival as a surrogate end-point for overall survival in patients with metastatic renal cell carcinoma. European Journal of Cancer, 50 (10), 1766-1771.

Özer-Stillman, I., Strand, L., Chang, J., Mohamed, A. F. and Tranbarger-Freier, K. E., 2015. Meta-analysis for the association between overall survival and progression-free survival in gastrointestinal stromal tumor. Clinical Cancer Research, 21 (2), 295-302.

Paoletti, X., Oba, K., Bang, Y.-J., Bleiberg, H., Boku, N., Bouché, O., Catalano, P., Fuse, N., Michiels, S., Moehler, M., Morita, S., Ohashi, Y., Ohtsu, A., Roth, A., Rougier, P., Sakamoto, J., Sargent, D., Sasako, M., Shitara, K., Thuss-Patience, P., Van Cutsem, E., Burzykowski, T. and Buyse, M., 2013. Progression-Free Survival as a Surrogate for Overall Survival in Advanced/Recurrent Gastric Cancer Trials: A Meta-Analysis. Journal of the National Cancer Institute, 105 (21), 1667-1670.

Petrelli, F. and Barni, S., 2013a. Correlation of progression-free and post-progression survival with overall survival in advanced colorectal cancer. Annals of oncology, 24 (1), 186-192.

Petrelli, F. and Barni, S., 2013b. Is overall survival still the primary endpoint in maintenance non-small cell lung cancer studies? An analysis of phase III randomised trials. Translational lung cancer research, 2 (1), 6.

Petrelli, F. and Barni, S., 2013c. Surrogate End Points and Postprogression Survival in Renal Cell Carcinoma: An Analysis of First-Line Trials With Targeted Therapies. Clinical Genitourinary Cancer, 11 (4), 385-389.

Petrelli, F. and Barni, S., 2014. Surrogate endpoints in metastatic breast cancer treated with targeted therapies: an analysis of the first-line phase III trials. Medical Oncology, 31 (1), 1-8.

Petrelli, F., Coinu, A., Borgonovo, K., Cabiddu, M. and Barni, S., 2015. Progression-free survival as surrogate endpoint in advanced pancreatic cancer: meta-analysis of 30 randomized first-line trials. Hepatobiliary & Pancreatic Diseases International, 14 (2), 124-131.

Pilz, R. L. and Manegold, C., 2013. Endpoints in lung cancer trials: today's challenges for clinical statistics. memo - Magazine of European Medical Oncology, 6 (2), 92-97.

Polley, M.-Y. C., Lamborn, K. R., Chang, S. M., Butowski, N., Clarke, J. L. and Prados, M., 2009. Six-month progression-free survival as an alternative primary efficacy endpoint to overall survival in newly diagnosed glioblastoma patients receiving temozolomide. Neuro-oncology, nop034.

Prasad, V., Kim, C., Burotto, M. and Vandross, A., 2015. The strength of association between surrogate end points and survival in oncology: A systematic review of trial-level meta-analyses. JAMA Internal Medicine, 175 (8), 1389-1398.

Prentice, R. L., 1989. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine, 8 (4), 431-440.

Sherrill, B., Amonkar, M., Wu, Y., Hirst, C., Stein, S., Walker, M. and Cuzick, J., 2008. Relationship between effects on time-to-disease progression and overall survival in studies of metastatic breast cancer. British journal of cancer, 99 (10), 1572-1578.

Sherrill, B., Kaye, J. A., Sandin, R., Cappelleri, J. C. and Chen, C., 2012. Review of meta-analyses evaluating surrogate endpoints for overall survival in oncology. OncoTargets and therapy, 5 287.

Shi, Q., De Gramont, A., Grothey, A., Zalcberg, J., Chibaudel, B., Schmoll, H.-J., Seymour, M. T., Adams, R., Saltz, L. and Goldberg, R. M., 2015. Individual patient data analysis of progression-free survival versus overall survival as a first-line end point for metastatic colorectal cancer in modern randomized trials: findings from the analysis and research in cancers of the digestive system database. Journal of Clinical Oncology, 33 (1), 22-28.

Shitara, K., Ikeda, J., Yokota, T., Takahari, D., Ura, T., Muro, K. and Matsuo, K., 2012. 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, 30 (3), 1224-1231.

Shitara, K., Matsuo, K., Muro, K., Doi, T. and Ohtsu, A., 2013. Progression-free survival and post-progression survival in patients with advanced gastric cancer treated with first-line chemotherapy. Journal of Cancer Research and Clinical Oncology, 139 (8), 1383-1389.

Shitara, K., Matsuo, K., Muro, K., Doi, T. and Ohtsu, A., 2014. Correlation between overall survival and other endpoints in clinical trials of second-line chemotherapy for patients with advanced gastric cancer. Gastric Cancer, 17 (2), 362-370.

Sidhu, R., Rong, A. and Dahlberg, S., 2013. Evaluation of progression-free survival as a surrogate endpoint for survival in chemotherapy and targeted agent metastatic colorectal cancer trials. Clinical Cancer Research, 19 (5), 969-976.

Singh, S., Wang, X. and Law, C., 2014. Association between time to disease progression end points and overall survival in patients with neuroendocrine tumors. Gastrointest Cancer Targets Ther, 4 103-113.

Suzuki, H., Hirashima, T., Okamoto, N., Yamadori, T., Tamiya, M., Morishita, N., Shiroyama, T., Takeoka, S., Osa, A., Azuma, Y. and Kawase, I., 2015. Relationship between progression-free survival and overall survival in patients with advanced non-small cell lung cancer treated with anticancer agents after first-line treatment failure. Asia-Pacific Journal of Clinical Oncology, 11 (2), 121-128.

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

Terashima, T., Yamashita, T., Takata, N., Nakagawa, H., Toyama, T., Arai, K., Kitamura, K., Yamashita, T., Sakai, Y. and Mizukoshi, E., 2015. Post-progression survival and progression-free survival in patients with advanced hepatocellular carcinoma treated by sorafenib. Hepatology Research.

Wilkerson, J. and Fojo, T., 2009. Progression-free survival is simply a measure of a drug's effect while administered and is not a surrogate for overall survival. The Cancer Journal, 15 (5), 379-385.

Yoshino, R., Imai, H., Mori, K., Takei, K., Tomizawa, M., Kaira, K., Yoshii, A., Tomizawa, Y., Saito, R. and Yamada, M., 2014. Surrogate endpoints for overall survival in advanced non-small-cell lung cancer patients with mutations of the epidermal growth factor receptor gene. Molecular and clinical oncology, 2 (5), 731-736.

Yoshino, R., Imai, H., Mori, K., Tomizawa, Y., Takei, K., Tomizawa, M., Kaira, K., Yoshii, A., Watanabe, S. and Saito, R., 2015. Clinical impact of postprogression survival for

overall survival in elderly patients (aged 75 years or older) with advanced nonsmall cell lung cancer. Journal of cancer research and therapeutics, 11 (3), 606.

5.2 References – Workshop, Technical Reviews and Conclusions

ABPI, 2013. Clinical Trial Transparency: Technical standards for data sharing for old, current and future clinical trials. Association of the British Pharmaceutical Industry (ABPI), Available: http://www.abpi.org.uk/industry-info/Documents/Clinical%20Trial%20Transparency.pdf.

Buyse, M., Burzykowski, T., Carroll, K., Michiels, S., Sargent, D. J., Miller, L. L., Elfring, G. L., Pignon, J.-P. and Piedbois, P., 2007. Progression-free survival is a surrogate for survival in advanced colorectal cancer. Journal of Clinical Oncology, 25 (33), 5218-5224.

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

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., der Elst, W. and Burzykowski, T., 2016. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. Biometrical Journal, 58 (1), 104-132.

Davis, S., Tappenden, P. and Cantrell, A., 2012. A review of studies examining the relationship between progression-free survival and overall survival in advanced or metastatic cancer. Sheffield: Decision Support Unit, ScHARR, University of Sheffield.

Department of Health (UK), 2010. Quantifying health impacts of government policies

Eisenhauer E.A, Therasse P., Bogaerts j. et al. 2009. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer, 45:228-247

European Medicines Agency (EMA), 2012. Guideline on the evaluation of anticancer medicinal products in man. London. EMA, Available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/01/WC500137128.pdf.

Griffin, S. C., Claxton, K. P., Palmer, S. J. and Sculpher, M. J., 2011. Dangerous omissions: the consequences of ignoring decision uncertainty. Health economics, 20 (2), 212-224.

Latimer NR, Abrams K, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, Akehurst RL, Campbell MJ. Adjusting survival time estimates to account for treatment switching in randomised controlled trials – an economic evaluation context: Methods, limitations and recommendations. Medical Decision Making 2014;34;3:387-402

Latimer NR. Survival analysis for economic evaluations alongside clinical trials – Extrapolation with patient-level data: Inconsistencies, limitations, and a practical guide. Medical Decision Making. 2013;33;6:743-754.

Mayor S. New "managed access" process for Cancer Drugs Fund to go ahead, NHS England confirms. BMJ 2016; 352.

Palmer, S. and Smith, P. C., 2000. Incorporating option values into the economic evaluation of health care technologies. Journal of health economics, 19 (5), 755-766.

Pilotto S, Carbognin L, Karachaliou N, Garassino M, Cuppone F, Petraglia S, et al. Moving towards a customized approach for drug development: lessons from clinical trials with immune checkpoint inhibitors in lung cancer. Transl Lung Cancer Res 2015;4:704–12.

Prentice, R. L., 1989. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine, 8 (4), 431-440.

Stevens, W., Philipson, T., Wu, Y., Chen, C. and Lakdawalla, D., A cost-benefit analysis of using evidence of effectiveness in terms of progression-free survival in making reimbursement decisions on new cancer therapies. Forum for Health Economics and Policy, 2014. 21-52.

US Food and Drug Administration (USFDA), 2007. Guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics.

ANNEX 1. DEFINITIONS OF DIFFERENT TYPES OF ENDPOINTS

(a) Overall survival (OS) and quality of life (QoL)

The most important outcome from a treatment for cancer is usually regarded as the additional time that an individual will live as a result of a particular treatment. For an unbiased measure of OS from a trial, it requires that OS must be measured for all trial participants. In some cases that may mean that the true OS does not become available for years or even decades.

For cost effectiveness analyses using quality adjusted life years (QALYs), mean length of survival is required, which again implies that an unbiased measure requires every person in the trial to have died. Many indicators of so-called OS, are in fact not of overall survival at all, but themselves are surrogate measures of OS. These include median survival, and 10-year, 5-year and 1-year survival percentages.

OS is not the only outcome of interest. QALY calculations and most other approaches to assessing value and value for money also require data on quality of life (QoL) between treatment and death. And in some forms and stages of cancer, this may also be important to patients' and clinicians' choice of treatment. So the "true" or ultimate endpoint may be a composite that includes OS and other characteristics such as QoL and patient preferences.

(b) Objective overall response rate (ORR)

This measure, the percentage of patients whose cancer has shrunk or disappears after treatment, was used until about 40 years ago before being largely supplanted by OS.

(c) Progression-free survival (PFS)

PFS (in an RCT context) is defined as the time elapsed between randomisation and tumour progression or death from any cause, whichever comes sooner. However, the definition of disease progression varies depending on the clinical trial and the kind of tumour analysed. For solid tumours, PFS normally measures the time from randomisation until the tumour has increased by 20% (in dimension, not volume) from the minimum to which it has shrunk due to treatment. RECIST criteria are commonly used (Eisenhauer et al., 2009).

(d) Progression on next-line therapy (PFS2)

This is defined as time from randomisation to objective disease progression on next-line therapy or death from any cause. Where PFS2 is used as an endpoint in oncology trials and survival benefit cannot be shown, there is a concern about changes of a tumour's drug resistance profile by experimental therapy. Currently, the literature on this is in its infancy and has not been included in the literature review.

(e) Time to progression (TTP)

This is similar to PFS except that it does not take account of deaths from other causes.

(f) Disease-free survival (DFS)

This is similar to PFS except that it ends at the time that the tumour has shrunk to its minimum size rather than the later point in time when the tumour has regained 20% of its size in each dimension.

(g) Time to treatment failure (TTF)

TTF is defined as the time from randomisation to treatment discontinuation for any reason, including disease progression, treatment toxicity, patient preference, or death. TTF is generally not accepted as a valid endpoint, as it is a composite endpoint influenced by factors unrelated to efficacy. Discontinuation may be a result of toxicity, patient preference, or a physician's reluctance to continue therapy. These factors are not a direct assessment of the effectiveness of a drug in shrinking tumours or impeding their growth.

(h) Post-progression survival (PPS)

In patients who have documented progression prior to death, PPS corresponds to the time from progression to death. It is commonly estimated as the difference between OS and PFS.

For fuller definitions please see US Food and Drug Administration (USFDA) (2007b).