Title: Methodological challenges for the evaluation of clinical effectiveness in the context of accelerated regulatory approval: an overview

Authors: Nerys Woolacott, Mark Corbett, Julie Jones-Diette, and Robert Hodgson, Centre for Reviews and Dissemination, University of York, UK. YO10 5DD.

Corresponding author: Nerys Woolacott, Centre for Reviews and Dissemination, University of York, UK. YO10 5DD Telephone +44 1904 321074 Email: nerys.woolacott@york.ac.uk

Date:

Funding: This report was commissioned by the NIHR HTA Programme as project number 14/151/06.

Declared competing interests of the authors: All authors have no personal or pecuniary conflict of interests.

Contributions of authors

All authors contributed to the planning and conduct of the review. NW and RH drafted the manuscript and MC and JJD commented on drafts. All Authors approved the final draft.

Abstract

Background

Regulatory authorities are approving innovative therapies with limited evidence. Whilst this level of data is sufficient for the regulator to establish an acceptable risk-benefit balance, it is problematic for downstream health technology assessment, where assessment of cost-effectiveness requires reliable estimates of effectiveness relative to existing clinical practice. Some key issues associated with a limited evidence base include using data, from non-randomised studies, from small single-arm trials, or from single-centre trials; and using surrogate endpoints.

Methods

We examined these methodological challenges through a pragmatic review of the available literature.

Results

Methods to adjust non-randomised studies for confounding are imperfect. The relative treatment effect generated from single-arm trials is uncertain and may be optimistic. Single-centre trial results may not be generalisable. Surrogate endpoints, on average, overestimate treatment effects. Current methods for analysing such data are limited and effectiveness claims based on these sub-optimal forms of evidence are likely to be subject to significant uncertainty.

Conclusions

Assessments of cost-effectiveness, based on the modelling of such data, are likely to be subject to considerable uncertainty. This uncertainty must not be underestimated by decision makers: methods for its quantification are required and schemes to protect payers from the cost of uncertainty should be implemented.

Word count – 200

Key words (6, American spelling)

**Accelerated approval; cost-effectiveness; non-randomised studies; single-arm trials; single-centre trials; surrogate outcomes**

What is new?
The limited data that are sufficient for a regulator to establish an acceptable risk-benefit balance can be problematic for cost-effectiveness analyses, which require reliable estimates of effectiveness relative to existing clinical practice.

A literature review of these methodological challenges identified that the main issues associated with a limited evidence base include using data from non-randomised studies, from small single-arm trials, or from single-centre trials; and using surrogate endpoints. Our review summarises, in a non-technical way, the principal methods adopted to deal with these issues and it highlights their limitations.

Assessments of cost-effectiveness based on the modelling of such data will be subject to huge uncertainty. This uncertainty must not be underestimated by decision makers: methods for its quantification are required and schemes to protect payers from the cost of uncertainty should be implemented.

# Background and objectives

In recent years, there have been increasing efforts to make innovative therapies available to patients more quickly. For example, during the last decade the European Medicines Agency (EMA) has been actively developing and introducing new regulatory approval pathway approaches including: approval under exceptional circumstances, conditional marketing authorisation, accelerated assessment, parallel scientific advice between EMA and the Food and Drug Administration (FDA) in the United States of America, and, most recently, the adaptive licensing pilot programme.1 Recently in the UK, an independently chaired report has been published, which sets out recommendations to speed up access to innovative healthcare and technologies and to improve efficiency and outcomes for NHS patients.2 For certain promising treatments, particularly the transformational ones referred to in the Accelerated Access Review, these initiatives may facilitate shorter approval times provided that adequate data on efficacy and safety are available, or will *become* available in the not too distant future. The result is that a growing number of promising, innovative new technologies will be granted EMA/FDA approval with limited or no data from randomised experiments. In such cases, the estimates of effectiveness will be based upon observational data, single-arm experimental studies and short-term studies with surrogate outcomes. Recent examples include Holoclar, which received EMA authorisation based on retrospective case series (combined n=148), and Glybera which was licensed based on single-arm studies (combined n=27). Crizotinib for 2nd line treatment of advanced anaplastic lymphoma kinase (ALK)-positive non-small cell lung cancer, received initial FDA approval based on objective response rate (ORR) and duration of response studied in a single-arm trial.3 A recent study of oncology drugs identified that a lack of randomised controlled trials (RCTs) was not a bar to approval by the FDA.4, 5

As regulatory authorities move to accept lower levels of evidence, there is a need to understand the limitations of such evidence, in the context of narrow populations with seemingly highly effective new treatments (high response rates). In particular, whilst this level of data is sufficient for the regulator to establish an acceptable risk-benefit balance, the availability of limited evidence is problematic for downstream health technology assessments, where assessment of cost-effectiveness requires reliable estimates of effectiveness relative to existing clinical practice.

As part of a larger project investigating the applicability of the National Institute for Health and Care Excellence (NICE) Technology Appraisal methods to the rapidly expanding field of regenerative medicines,6 we examined these methodological challenges through a review of the available literature. The focus of this paper is to highlight the issues faced when interpreting and using these types of data. Specifically this paper seeks to address the following questions and to present an overview of the current position:

* What are the main issues with evidence from non-randomised studies (NRS)? Specifically
* To what extent do estimates of effectiveness, obtained from NRS, agree with those obtained from randomised trials (i.e. is there any reasonable quantification of bias that can be applied to NRS?). What techniques are available to adjust for confounding in NRS and how reliable are they?
* What are the specific challenges of using single-arm studies to estimate treatment effectiveness?
* What issues may arise when interpreting the results of single-centre trials, compared with multi-centre trials?
* What are the issues associated with using surrogate endpoints as the primary outcome measure in trials of therapeutic agents?

# Methods

We conducted pragmatic reviews of the literature to address these research questions. These reviews were conducted between May and December 2015. In our reviews, the term non-randomised studies (NRS) encompasses non-randomised trials including controlled trials and single-arm trials, and controlled and uncontrolled observational studies. For each review we started with studies already known to the team and identified further key articles using unstructured searches of MEDLINE. Specifically, the review of bias in NRS started with available HTA monographs, 7, 8 and the review of surrogate endpoints began with a search of key guidelines on the use of surrogate endpoints produced by the FDA,9 NICE Decision Support Unit (DSU: University of Sheffield),10 EUnetHTA11 and survey results produced by the NIHR HTA Programme of reports on the cost-effective use of surrogate outcomes.12 Based on these key studies snowballing techniques were then applied; citation searches were carried out and further unique references checked for relevant studies. Citations and references of any additional studies identified were then also checked until no further relevant studies were identified.

The screening of records for inclusion in the review was not conducted in duplicate. Data extraction and the synthesis of identified relevant information were checked by a second reviewer.

# Results

## Quantification of bias and adjustment for confounding in non-randomised studies

A total of 14 studies7, 8, 13-24 were identified as relevant to the quantification of bias in NRS (details are presented in the full technical report).6 In most of these studies the quantification of bias was estimated by directly comparing results from otherwise similar NRS and randomised studies: the evidence from randomised and non-randomised sources was pooled separately and the summary effects were compared. The comparison was based on outcome measures such as: assessment of direction of effect; subjective assessment of overlap of confidence intervals, and proximity of summary estimates; tests of statistical difference in summary estimates of effect obtained from randomised and non-randomised evidence; and the calculation of ratios of odds or risk ratios. In many studies, multiple outcomes were used. The lack of consensus around how to measure the degree of concordance between results obtained from randomised and non-randomised studies makes it difficult to interpret the results from individual studies and presents a significant barrier to making comparisons across studies.

Of the 14 included studies, seven7, 8, 14, 15, 17, 19, 22 concluded that there were no systematic differences in either the size or direction of effect estimates obtained from NRSs compared with those from RCTs. Five studies (including the largest study that contained RCTs and NRSs from 45 topic areas21) concluded that effect estimates obtained from NRSs were systematically larger than those obtained from RCTs.13, 16, 18, 21, 23 The authors of the other two studies20, 24 were unable to draw any meaningful conclusions about the comparability of their estimates.

Study design was identified as a factor in determining the reliability of estimates of clinical effectiveness obtained from NRSs. Empirical evidence of the potential for bias in studies using a historical control group was investigated in three studies that all found greater concordance between the results of RCTs and NRSs when studies with historical controls were excluded.14, 19, 21 The largest study,21 found that compared with retrospective NRSs, the results from prospective NRSs (either with current or historical controls) were closer to those from randomised studies. In this context it is worth noting that single-arm trials use an implied historical control. An investigation into broader measures of quality revealed that there was a tendency towards smaller discrepancies between high-quality NRSs and RCTs than between low-quality NRSs and RCTs.8

The impact of bias can be ameliorated by adjusting for confounding. These methods have been summarised in a number of articles, including review articles25-27 and a recent DSU report28 of methods of adjustment in NRS. These methods can be divided into those that adjust for bias at the study level,26, 27, 29-35 and those that adjust for bias in the evidence synthesis process.36-44 These are summarised in Table 1. Our review of studies comparing the relative effectiveness of different methods of adjustment identified nine studies, which are summarised in Table 2.

Table 1 Summary of adjustment methods for non-randomised studies

|  |  |
| --- | --- |
| Method | Brief description |
| **Adjusting for bias at the study level** |
| Regression analysis | Seeks to directly adjust for these potential confounding variables by building a statistical model27 of the form: Outcome variable = f (control variables + treatment decision)Regression models allow the estimation of the treatment effect conditional upon these confounding variables. |
| Propensity scoring  | Propensity scoring rather than being a single method is a suite of methods that consider confounding bias as a form of selection bias where treatment allocation is acknowledged to be non-random and that treatment selection is often influenced by a patient’s characteristics.26 The various methods are: matching, stratification, inverse probability of treatment weighting (IPTW), and covariate. |
| Instrumental variables | Instrumental variables techniques attempt to approximate the experimental approach by using an instrument variable or variables. A parameter is considered a valid instrument if it meets the following two conditions; 1) The instrument must be correlated with the receiving of treatment (or exposure); 2) The instrumental variable must be independent (uncorrelated) with unobserved confounders. |
| Stratification | The division of participants into subgroups with respect to categorical (or categorised quantitative) prognostic factors, for example, classifying age into decades, or weight into quartiles. The intervention effect is then estimated in each stratum, and a pooled estimate is calculated across strata. This procedure can be interpreted as a meta-analysis at the level of an individual study. 27 |
| **Matching** | Selecting participants with similar values for important prognostic factors to make the control and treatment groups more similar, and to prevent any differences between the treatment and control group being as a result of differences in the matched variables. Matching can be carried out either prospectively or retrospectively. It is unlikely to completely account for all differences due to unobserved confounding.32 |
| **Suppression of bias factors in single arm trials** | Suppression of four bias factors: attrition bias, bias from natural recovery, regression to the mean, and bias from adjunctive therapies. 33  |
| **Adjusting for bias at the evidence synthesis level** |
| **Adjusting using external data** | A Bayesian hierarchical model to model bias in RCTs that are at high risk of bias. In the mixed-effects model, treatment effects are considered as fixed and bias effect as random. Estimates of bias in any given meta-analysis are given as a function of prior distribution, which is estimated from published meta-analysis of RCTs, and data from the current meta-analysis.36-42 |
| **Elicitation** | The direction and magnitude of biases are elicited by reviewers. 43, 44 |

Table 2 Studies comparing bias adjustment methods

|  |  |  |
| --- | --- | --- |
| **Study name and design** | **Methods compared**  | **Summary of findings** |
| Biondi-Zoccai (2011)45 Discussion piece | Regression analysis, Propensity scoring and instrumental variables | Propensity scoring may have advantages over other methods of adjustment, but all methods have important limitations. |
| Cepeda (2011)46 Simulations study | Propensity scoring and logistic regression | Logistic regression is superior to propensity scoring when the number of events is greater than 8 per confounder |
| Crosby (2010)47Comparison | Regression analysis and instrumental variables | Note some discrepancies in results obtained using regression analysis and instrumental variables. Suggest that instrumental variables may be superior to regression analysis as a method of accounting for confounding bias.  |
| Kurth (2006)48Comparison | Propensity scoring and logistic regression | Different methods to control for confounding yielded extremely different treatment effect estimates. This disparity was suggested to be a result of each analysis answering a different question implicit or explicit to that method of adjustment. |
| Laborde-Casterot 201549 Systematic review | Propensity scoring with instrumental variable analyses | There was slight/fair agreement between the methods [Cohen’s kappa coefficient = 0.21 (95% CI 0.00-0.41)]. In 42% of cases the two methods produced different results in terms of results being significant / non-significant; using instrumental variable methods results were non-significant in 87% of cases. |
| Martens (2008)50 Simulation study | Propensity scoring and logistic regression | On average, estimates from propensity scoring are closer to the true marginal treatment effect than those generated by logistic regression.  |
| Shah (2005)51 Systematic review | Propensity scoring and standard regression analysis | Observational studies had similar results whether using traditional regression or propensity scores to adjust for confounding. Propensity scoring produced on average more conservative estimates of effect. |
| Stukel (2007)52Comparison | Regression analysis, Propensity scoring (via matching and covariate adjustment) and instrumental variables | Estimates of the observational association of cardiac catheterisation with long-term AMI mortality were highly sensitive to the analytic method. Compared with standard modelling, instrumental variable analysis may produce less biased estimates of treatment effects. |
| Sturmer (2006)53Comparison | Propensity scoring and standard regression analysis | Little evidence that these methods yielded substantially different estimates compared with conventional multivariable methods. |

Comparisons of alternative methods of adjustment reveal some inconsistency in the results obtained from the different methods of adjustment, but do not indicate the superiority of one method of adjustment over another. Overall, it is unclear which methods are most appropriate in any given circumstance, and further research is needed. Furthermore, adjusting for confounding often requires individual patient data (IPD), which can be difficult to access. Approaches for recreating IPD have been developed, such as the algorithm by Guyot et al.54

In summary, in many instances NRS do generate results that are different from those of an equivalent RCT, although quantifying the effects of the various biases that NRSs are subject to is difficult. The performance of adjustment methods to eliminate confounding biases remains uncertain and, as such, the results from NRS will be subject to a degree of uncertainty, even after adjustment for confounding.

## Challenges of using single-arm trials to estimate effectiveness

Ten articles were identified as relevant to the issue of using single-arm trials to estimate effectiveness.55-65 One of these was a review paper55 on the opportunities and challenges in using data from uncontrolled studies.

Despite the lack of a direct, concurrent comparator in single group studies, both explicit and implicit comparisons are frequently made. Implicit comparisons are made when the expected outcomes in the absence of the intervention of interest, are believed to be well known, and the expected effect size from the intervention is large. Explicit comparisons are made when the investigators compare the single group of subjects before and after an intervention, or when the investigators choose to incorporate a specified historical comparator in the analysis. Each of these alternative study designs has particular challenges and advantages.55 The particular challenges are discussed below.

### Implicit comparisons

Implicit comparisons are common: a survey found that roughly half of Phase II studies did not cite the source of their historical response rates.56 Examples of implicit comparisons can be seen in the EMA approved drugs for hepatitis C, recently appraised by NICE (sofusbuvir, daclatasvir, ledipasvir and simeprevir),66 where because of the objective outcome, the known lack of recovery without treatment, and the large treatment benefit, regulatory approval had been granted based on short-term single-arm trials. Even in such cases there may be clinically relevant differences between the patients who are enrolled in the single–arm trial and those who do not qualify. Careful review of the study population and eligibility criteria is needed to make an assessment concerning external validity in relation to the implied control.55 An important limitation of implicit comparisons is that they cannot be used in cost-effectiveness analysis, as this requires an explicit comparison; in the above example appraisal by NICE of sofusbuvir, daclatasvir, ledipasvir and simeprevir for hepatitis C, explicit comparisons were made, based on historical control groups, to establish cost-effectiveness.

### Explicit comparisons

##### Before-and-after designs

Before-and-after designs (pre-post designs) assess the difference in response before and after the administration of an intervention, in a single group of patients: patients serve as their own controls. To provide unbiased estimates of effectiveness from such studies it is necessary to eliminate all alternative explanations for the observed treatment effects such as: adjunctive therapy; carryover effects from previous therapies; natural recovery; or the effects of regression to the mean.57 Therefore, before-and-after designs are only appropriate for stable chronic conditions where any variation in disease status is highly unlikely without the effect of an intervention.

##### Historical controls

Comparative estimates of effectiveness can be generated from single-arm studies by comparing the study results with relevant historical data. The validity of such historical comparisons is limited by differences between the historical patients and those in the single-arm study. Differences in known and unknown prognostic factors, or changes in diagnostic methods or outcome ascertainment are important. Furthermore, the selection of the historical control may be subject to bias: for new agents there is a natural enthusiasm amongst the investigators for the new agent and this may manifest itself in the selection of a low historical response rate (from a real or implied historical cohort).56

To address the problem of reliable historical controls for single-arm trials, efforts have been made to amass historical databases and derive historical control data for future trials in specific disease sites. Examples include stage IV melanoma, advanced pancreatic cancer58 and advanced non–small cell lung cancer.59 However, in such rapidly evolving therapeutic areas the continued relevance of any historical cohort should always be questioned. For example the NICE technology appraisal (TA410) talimogene laherparepvec for treating unresectable metastatic melanoma, the Korn data set was rejected as being no longer relevant because of newly available immunotherapy treatment options (to which patients in the Korn cohort had not been exposed; <https://www.nice.org.uk/guidance/ta410>).

##### Comparability of results from single-arm studies and randomised designs

In the context of the question of whether Phase II trials should be single-arm or randomised, a number of studies have sought to compare the estimates of effectiveness from single-arm studies with those from randomised trials.61-64 These simulation studies all indicated that when the treatment effect of the historical control is a known quantity and does not vary, then the results from single-arm trials reflect those from randomised trials, correctly identifying a real treatment benefit. However, in reality there is always uncertainty around any presumed historical control estimate, and the simulations demonstrated that even a modest variation in the historical control treatment effect greatly increases the ‘false-positive’ rate of single-arm trial results. Furthermore, increasing the size of the single-arm trial is counter-productive as it inflates, rather than reduces, the false positive rate.61-64 This is not unexpected as increasing the sample size does nothing to eliminate bias and only serves to provide a more precise, but still biased estimate of effect. For the results from single-arm studies to be used fully in HTA, estimates of the uncertainty of the treatment effects from empirical studies are required.

## Generalisability of single-centre trials

It has been suggested that single-centre trials may produce significantly larger effect estimates than multi-centre trials.67-69 Possible reasons for the larger effects are that single-centre studies: recruit fewer patients than multi-centre studies (smaller studies tend to report larger effects); may have larger treatment effects due to the high levels of centre expertise; and may recruit populations which are unduly homogenous. Also, if the single-centre is a specialist academic medical centre, the patients may be atypical, in that they may be more mobile, more heavily pre-treated, have better socio-economic status, or receive better supportive care.60 A meta-epidemiological study, based on 119 core clinical journals and the Cochrane Database of Systematic Reviews, for meta-analyses of RCTs , published during 2012 that included 241 RCTs (binary outcome) and 173 RCTs (continuous outcome), found no difference in magnitude of effect between single-centre and multi-centre trials for binary outcomes (ratio of odds ratios (ROR) 1.02 [95% CI: 0.83, 1.24]), but effect sizes were systematically larger for single-centre than for multi-centre trials for continuous outcome measures (mean difference in standardised mean difference (SMD)s: −0.13 [95% CI: −0.21 to−0.05]).70

It is, however, worth acknowledging here that multi-centre trials where the recruitment of patients is unnecessarily dispersed across a large number of centres (marketing trials) may also recruit atypical patients or have other methodological flaws due to a lack of trial-based expertise in the numerous centres.71

## Challenges with the use of surrogate endpoints as primary outcome measures in definitive effectiveness trials of new therapeutic agents

Our review of issues associated with using surrogate endpoints as the primary outcome measure in trials of new therapeutic agents identified 33 articles (details are presented in the full technical report)6. There was significant overlap between these articles and they are summarised in the following narrative.

Ideally, the relative effectiveness of drugs and treatments should be based on final clinical endpoints.11, 72 However, randomised controlled trials with large sample sizes and extended follow-up periods are required to capture the statistical significance of a treatment or intervention’s impact on such an outcome.11 In contrast, the use of a surrogate outcome can enable a more rapid assimilation of data, which can be important for new therapies for patients, where mortality is high, morbidity is debilitating or treatment options are few.12 Issues with the use of surrogate outcomes have been discussed in the literature for many years: whilst there are conditions under which surrogate outcomes are appropriate, in reality such conditions are rarely met and there are doubts over the validity of surrogate outcomes and concerns about the risks and potential adverse consequences, associated with their use in clinical research.73-80 The results of a recent questionnaire study, of 74 stakeholders in the development of drugs for cardio-renal disease, indicated that although the use of surrogates is not opposed, most are not considered valid: out of the four surrogate outcomes suggested as an endpoint for trials - blood pressure, glycated haemoglobin, albuminuria or C-reactive protein - only blood pressure was considered moderately accurate.81

Overall survival (OS) is considered to be the gold standard to measure benefit in many clinical trials, particularly oncology trials, as it provides a precise, statistically and clinically meaningful endpoint. It is also a key parameter in health economic models, where it is often the key driver of cost-effectiveness. However, mature OS data are difficult to acquire due to the length of follow-up needed and the number of deaths required for appropriate statistical analyses. This is true even in studies of terminal illnesses such as metastatic cancer. As a solution, there has recently been a steady move (by regulatory bodies) away from OS, as a clinical endpoint measure, and towards more short-term surrogate measures, such as progression-free survival (PFS) or even objective response rate (ORR), particularly in the context of accelerated approvals or where there is a high unmet patient need. 82-84

From a regulatory and clinical HTA perspective, the absence of data on clinical endpoints might be acceptable when a clinical endpoint is difficult or impossible to study, e.g. when populations are rare and there is a high unmet clinical need.9, 11 Recommendations for the use of surrogates in HTA have been made:11, 12, 72 efficacy assessments of pharmaceuticals should be based whenever possible on final patient-relevant clinical endpoints (e.g. morbidity or, overall mortality), but where a surrogate is used the choice of surrogate outcomes must be researched, explicit and justified. Ideally a systematic review of the evidence for the validation of the surrogate/final outcome relationship should be performed and the evidence on surrogate validation should be presented according to an explicit hierarchy.

There is significant literature on the issue of the validation of surrogates, and there has for some time been a number of guidelines proposed for assessing validity.11, 12, 79, 85 Further work has also been published on scoring schemas for the value of surrogates86 and methods for the statistical validation of surrogates as outcome measures have been developed.87-90 In recent reviews of current statistical approaches to surrogate endpoint validation, based on meta-analysis in various advanced-tumour settings, the suitability of surrogates for OS, mainly ORR, PFS and time-to-progression (TTP), have been examined.10, 91-94 The findings suggest that the strength of the association between the surrogates and OS is generally low and does not necessarily hold across cancer types, though a patient-level analysis indicated that responders to treatment have longer PFS and OS than do non-responders.94

A recently published analysis of the degree of difference in treatment effects between PFS and TTP and OS in RCTs of pharmacologic therapies in advanced colorectal cancer found a larger treatment effect for the surrogates than for OS.93 Across a broader range of surrogate outcomes a meta-epidemiological study of 185 trials found that on average, trials using surrogate outcomes (n=84) reported treatment effects that were 28% to 48% higher than those using final patient relevant outcomes (n=101), and this result was consistent across sensitivity and secondary analyses.95

Given the difficulty in validating surrogate outcomes, which conflicts with their use in clinical research, Ciani and Taylor 96 comment that the need for pragmatic high level evidence, preferably from meta-analyses and regression modelling using both surrogate and final outcomes for HTA, must be more widely recognised. The potential of this is demonstrated by a study conducted to illustrate the potential to reduce uncertainty around the clinical outcome by estimating it from a multivariate meta-analysis.97 By allowing all the relevant data to be incorporated in estimating clinical effectiveness outcomes - including data from surrogate outcomes - multivariate meta-analysis can improve the estimation of health utilities through mapping methods.

In the context of HTA the use of surrogate outcomes, such as ORR or PFS, as primary endpoints in clinical trials mean that the OS data are never mature, making decision modelling difficult and the results highly uncertain, e.g. TA406 Crizotinib for untreated ALK-positive advanced non-small cell lung cancer (<https://www.nice.org.uk/guidance/ta406>).

# Discussion

A desire to get innovative promising medicines to market quickly means that accelerated regulatory approvals are likely to be supported by less than ideal data. In this paper, we consider the main issues regarding the type of data used to obtain early regulatory approval: non-randomised comparisons; single-arm studies; single-centre trials; and surrogate outcomes. It is important that the research community has a clear awareness of the challenges associated with data that are sub-optimal because of these methodological issues. Whilst there are few effective options available to address them, our findings provide a useful context to work carried out by the NICE DSU28, 98 and others, which provides a technical overview of methods for analysing NRS and other types of sub-optimal data. Specifically, our review presents a non-technical summary and highlights the point that, despite recent technical advances in a number of methods of analysis, significant challenges remain when using suboptimal data to estimate treatment effectiveness.

Whilst we believe that the issues summarised in this article are the most important ones for HTA, we acknowledge that there are additional issues, not covered in this paper, which may be of relevance for HTA in certain contexts. Examples of important issues include the growing use of physician's choice or blended comparators, as the control treatment in oncology trials, and the use of adaptive trial designs (e.g., basket trials, umbrella trials) that are emerging in oncology.99-101 Furthermore, although our review was conducted systematically, it was not comprehensive nor exhaustive. The methods used enabled us to be confident that the key methodological studies or reviews, and sufficient examples related to the relevant issues under discussion were captured, but not that every individual study or potentially relevant article was reviewed.

For NRS the challenges centre on when and if estimates of treatment effectiveness obtained from NRS can be relied upon, and much of the literature to date has focused on quantifying the bias in NRS. The current literature indicates that the degree of discordance between the findings from NRS and RCTs depends upon the specific study design used (being prospective is very important) and overall study quality. The literature further indicates that whilst methods of analysis to adjust for bias have been developed, it is unclear which methods are most appropriate in any given circumstance and further research is needed. Consequently, while there are individual examples in which RCTs and NRS produce comparable results, the risk of confounding remains and estimates generated from NRS will be subject to a degree of uncertainty, even after adjustment for confounding.

Related to the use of NRS is the role single-arm studies have in establishing effectiveness. Our review identified some simulation studies that suggest that single-arm trials can only be reliable indicators of treatment benefit when the natural history of disease is a known quantity, the patient population is homogenous, and the control treatment has little impact on outcomes, and when such conditions are not met, increasing the size of single-arm trials may not be helpful. Given how unlikely it is that such conditions are met in reality, HTA requires strong empirical estimates of the uncertainty around the results from single-arm trials. The availability of historical databases, from which to derive historical control data for future trials in specific disease sites, is extremely important for better evaluation and analysis of data from single-arm trials, and is essential to generate the estimates of relative effectiveness needed in economic models for the assessment of cost-effectiveness. However, in rapidly evolving therapeutic areas, the continued relevance of any such historical cohort should always be questioned. In certain specific circumstances, it may be appropriate to use data from single-arm studies, which, potential biases notwithstanding, may be more reliable than from the available RCTs, for example when the RCT evidence is limited to very small-scale, single-centre trials. The issue with single-centre trials, is that even if they have good internal validity, there is a risk that the results will not be generalisable to broad clinical practice.

Surrogate outcomes present an opportunity to radically improve the efficiency of technology development by allowing data collection to focus on outcomes that, while less clinically relevant are easier and quicker to achieve. There are clear recommendations regarding the choice, validation and use of surrogate outcomes, but the problem of whether the outcome is a truly reliable surrogate for the final clinical outcome often remains. Studies looking at surrogates for overall survival demonstrate how difficult it is to validate even commonly used surrogates, such as PFS. Where surrogate outcomes are accepted for regulatory purposes there is still the huge challenge of finding a reliable estimate of the real clinical benefit to populate the economic model for health technology assessment. One pragmatic suggestion is that analyses, at whatever stage of development and maturity of data, should include all the available outcome data in order to minimise uncertainty.96

With respect to the methodological issues addressed in this article, challenges clearly remain and current methods are not in a position to “fix” sub-optimal data. Decisions made using such data are inevitably subject to greater uncertainty than those made with more optimal data. This added uncertainty will generate challenges for regulators assessing the benefits and risks of new therapies. These risks may be offset by more rigorous implementation of conditional approval processes and enforcement by the regulator of demands for ongoing data collection and assessment. The existence of such uncertainty can perhaps be justified for innovative medicines targeted at tightly defined conditions, with a narrow population to minimise heterogeneity, where otherwise patients have little or no chance of recovery or improvement. This is supported by the recommendations in the recent Accelerated Access Report.102

The availability of such limited data, however, presents challenges and problems for HTA and economic modelling, where a reliable estimate of the treatment effect relative to best supportive care or the appropriate active comparator, is essential. Inevitably the cost-effectiveness of new technologies in these circumstances will be more difficult to determine and may produce estimates of cost-effectiveness subject to potentially unacceptable levels of uncertainty. One such example is ceritinib for 2nd- or 3rd-line non-small-cell lung cancer. Ceritinib received EMA approval for patients with ALK-positive non‑small‑cell lung cancer, who had been treated previously with crizotinib. The clinical evidence base for this approval comprised two single-arm trials, although they did provide mature progression free and overall survival data. For the purposes of the NICE appraisal (TA395),103 historical control (best supportive care) data were sourced from two separate observational studies: one each for PFS and OS. In this case, the NICE appraisal committee concluded that ceritinib was likely to prolong life, but the extent of treatment benefit was highly uncertain. In the context of this uncertainty, they concluded that the benefit was likely to meet their minimum threshold, but called for more robust evidence for future appraisals. The impact of NICE’s positive recommendation, and the risk associated with an incorrect decision, was limited by recommending ceritinib for only the small number of patients who had previously received crizotinib and by the price discount offered by the company (via a Patient Access Scheme).

For reimbursement agencies to be able to routinely assess interventions with sub-optimal data there is a need for these agencies to adapt their methods of assessment and for them to consider schemes that allow the development of further evidence, possibly with a risk-sharing component. Managed entry agreements (MEAs) are an increasingly common policy response to dealing with uncertainty in the evidence for new health technologies entering the market. These include performance-based risk sharing agreements, where the level or continuation of reimbursement is based on the health and economic outcomes achieved. These would allow the risk to be distributed appropriately across both the NHS and the manufacturers, while allowing reimbursement agencies to continue to support patient access to new and innovative therapies. Such schemes, however, have their limitations and determining where MEAs are an appropriate policy instrument remains a key methodological challenge with implications for both policy makers and manufacturers. The potential for the use of MEA’s was explored theoretically in a modelling exercise, presented in our report, and issues relating to their application were explored (see full HTA report for details).6

The increasing use of sub-optimal data for regulatory approval demands that a substantial research effort is applied to the methods and procedures for handling sub-optimal data to help decision makers in the context of such uncertainty. This should include research into the effectiveness of alternative methods to adjust for bias in NRS and further research into appropriate methods for validating surrogate outcomes, as well as the establishment of best practice guidelines to support both the development and assessment of submissions. Most urgently, research is needed to quantify the uncertainty generated by the use of sub-optimal data and its impact on decision making. Only by understanding the impact of uncertainty on decision making is it possible to accurately determine the value of further research and therefore formally cost the use of sub-optimal data in reimbursement decisions.

# Conclusions

Accelerated regulatory approval pathways, whereby treatments receive approval based on limited data, present a challenge for HTA. The methods of HTA and economic evaluation depend upon reliable estimates of relative effect for final clinical outcomes. Whilst there is an awareness of the issues, and methods have been proposed to adjust for bias in non-randomised studies, and to adjust for the lack of a suitable comparator, and for the use of surrogate outcomes, results based on the modelling of such data will be subject to huge uncertainty. This uncertainty must not be underestimated by decision makers: methods for its quantification are required and schemes to protect payers from the cost of uncertainty should be implemented.

Acknowledgements

We acknowledge and thank Kath Wright for her help in searching for studies and managing the references.

This report was commissioned and funded by the NIHR HTA Programme as project number 14/151/06. The views expressed in this report are those of the authors and not necessarily those of the NIHR HTA Programme. Any errors are the responsibility of the authors.

# References

1. Baird LG, Banken R, Eichler HG, Kristensen FB, Lee DK, Lim JC, et al. Accelerated access to innovative medicines for patients in need. *Clin Pharmacol Ther* 2014;96:559-71.

2. Department of Health. *Accelerated access review: final report. Review of innovative medicines and medical technologies*; 2016. Available from: <https://www.gov.uk/government/publications/accelerated-access-review-final-report>

3. Selaru P, Tang Y, Huang B, Polli A, Wilner K, Donnelly E, et al. Sufficiency of single-arm studies to suport registrtaion of targeted agents in molecularly selected patients with cancer: lessons from the clincial development of crizotinib. *Clinical Translational Science* 2016;00:1-11.

4. Tibau A, Ocana A, Anguera G, Seruga B, Templeton A, Barnadas A, et al. Oncologic Drugs Advisory Committee Recommendations and approval of cancer drugs by the US Food and Drug Administration. *JAMA Oncol* 2016;2:744-50.

5. Tsimberidou A, Braiteh F, Stewart D, Kurzrock R. Ultimate fate of oncology drugs approved by the US Food and Drug Administration without a randomized trial. *J Clin Oncol* 2009;27:6243-50.

6. Hettle R, Corbett M, Hinde S, Hodgson R, Jones-Diette J, Woolacott N, et al. The assessment and appraisal of regenerative medicines and cell therapy products: an exploration of methods for review, economic evaluation and appraisal. *Health Technol Assess* 2017;21:1-204.

7. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;2:1-136.

8. MacLehose R, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;4:1-162.

9. Katz R. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx: the American Society for Experimental NeuroTherapeutics* 2004;1:189–95.

10. Davis S, Tappenden P, Cantrell A. *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; 2012.

11. EUnetHTA. *Endpoints used in relative effectiveness assessment of pharmaceuticals: surrogate endpoints*; 2013. Available from: <http://www.eunethta.eu/sites/5026.fedimbo.belgium.be/files/Surrogate%20Endpoints.pdf>

12. Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. *Health Technol Assess* 2009;13:88.

13. Abraham NS, Durairaj R, Young JM, Young CJ, Solomon MJ. How does an historic control study of a surgical procedure compare with the "gold standard"? *Dis Colon Rectum* 2006;49:1141-8.

14. Algra AM, Rothwell PM. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *Lancet Oncol* 2012;13:518-27.

15. Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. *N Engl J Med* 2000;342:1878-86.

16. Beynon R, Savovic J, Harris R, Altman D, Sterne J, Deeks J. Quantification of bias in the results of non-randomised studies compared with randomised studies. *Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen* 2008;102:69.

17. Concato J, Shah N, Horwitz RI. Randomized controlled trails observational studies and the hierarchy of research designs. *N Engl J Med* 2000;342:1887-92.

18. Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J* 2012;33:1893-901.

19. Golder S, Loke YK, Bland M. Comparison of Pooled Risk Estimates for Adverse Effects from Different Observational Study Designs: Methodological Overview. *Plos One* 2013;8.

20. Hartz A, Bentler S, Charlton M, Lanska D, Butani Y, Soomro GM, et al. Assessing observational studies of medical treatments. *Emerg Themes Epidemiol* 2005;2:8.

21. Ioannidis JPA, Haidich A-B, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821-30.

22. Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg* 2014;259:18-25.

23. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233-40.

24. Shepherd J, Bagnall A-M, Colquitt J, Jacqueline D, Duffy S, Sowden A. *Sometimes similar, sometimes different: a systematic review of meta-analyses of randomised and non-randomised policy intervention studies*: Wessex Insitute for Health Research and Development. University of Southampton. Centre for Reviews and Dissemination. University of York; 2006.

25. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods* 2015;6:45-62.

26. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res* 2011;46:399-424.

27. McNamee R. Regression modelling and other methods to control confounding. *Occup Environ Med* 2005;62:500-06.

28. Faria R, Hernandez Alava M, Manca A, Wailoo AJ. *The use of observational data to inform estimates to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data*. Sheffield: Decision Support Unit. ScHARR. University of Sheffield; 2015.

29. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *Int J Biostat* 2009;5:Article 13.

30. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* 2007;26:734-53.

31. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006;25:2084-106.

32. Bland JM, Altman DG. Matching. 9. *BMJ* 1994;309:1128-28.

33. Hamre HJ, Glockmann A, Kienle GS, Kiene H. Combined bias suppression in single-arm therapy studies. *J Eval Clin Pract* 2008;14:923-9.

34. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York ; London: Springer; 2001.

35. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004;23:2937-60.

36. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. *J R Stat Soc Ser A* 2009;172:119-36.

37. Doi SA. Evidence synthesis for medical decision making and the appropriate use of quality scores. *Clin Med Res* 2014;12:40-6.

38. Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *Int J Technol Assess Health Care* 1993;9:440-9.

39. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Stat Med* 2006;25:2136-59.

40. Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J R Stat Soc a Stat* 2008;171:159-78.

41. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Stat Med* 2000;19:3359-76.

42. Soares MO, Dumville JC, Ades AE, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. *J R Stat Soc a Stat* 2014;177:259-79.

43. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat Med* 2003;22:3687-709.

44. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. *J R Stat Soc Ser A* 2009;172:21-47.

45. Biondi-Zoccai G, Romagnoli E, Agostoni P, Capodanno D, Castagno D, D'Ascenzo F, et al. Are propensity scores really superior to standard multivariable analysis? *Contemp Clin Trials* 2011;32:731-40.

46. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158:280-7.

47. Crosby DA, Dowsett CJ, Gennetian LA, Huston AC. A tale of two methods: comparing regression and instrumental variables estimates of the effects of preschool child care type on the subsequent externalizing behavior of children in low-income families. *Dev Psychol* 2010;46:1030-48.

48. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol* 2006;163:262-70.

49. Laborde-Castérot H, Agrinier N, Thilly N. Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. *J Clin Epidemiol* 2015;68:1232-40.

50. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *Int J Epidemiol* 2008;37:1142-7.

51. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005;58:550-9.

52. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;297:278-85.

53. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;59:437-47.

54. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012;12:9.

55. Paulus JK, Dahabreh IJ, Balk EM, Avendano EE, Lau J, Ip S. Opportunities and challenges in using studies without a control group in comparative effectiveness reviews. *Res Synth Methods* 2014;5:152-61.

56. Vickers A, Ballen V, Scher H. Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. *Clin Cancer Res* 2007;13:972-6.

57. Torgerson DJ, Torgerson C. *Designing randomised trials in health, education and the social sciences : an introduction*. Basingstoke, Hampshire: Palgrave Macmillan; 2008.

58. Korn E, Liu P, Lee S, Chapman J, Niedzwiecki D, Suman V, et al. Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials. *J Clin Oncol* 2008;26:527-34.

59. Philip P, Chansky K, LeBlanc M, Rubinstein L, Seymour L, Ivy S, et al. Historical controls for metastatic pancreatic cancer: benchmarks for planning and analyzing single-arm phase II trials. *Clin Cancer Res* 2014;20:4176-85.

60. Sargent D, Taylor J. Current issues in oncology drug development, with a focus on Phase II trials. *J Biopharm Stat* 2009;19:556-62.

61. Tang H, Foster NR, Grothey A, Ansell SM, Sargent DJ. Excessive false-positive errors in single-arm phase II trails: a simulation-based analysis. *J Clin Oncol* 2009;25 Suppl. 15:abstr 6512.

62. Tang H, Foster N, Grothey A, Ansell S, Goldberg R, Sargent D. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol* 2010;28:1936-41.

63. Pond GR, Abbasia S. Quantitative evaluation of single-arm versus randomized phase II cancer clinical trials. *Clin Trial* 2011;8:260-69.

64. Sambucini V. Comparison of single-arm vs. randomized phase II clinical trials: a Bayesian approach. *J Biopharm Stat* 2015;25:474-89.

65. Monzon J, Hay A, McDonald G, Pater J, Meyer R, Chen E, et al. Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome. *Eur J Cancer* 2015;51:2501-07.

66. National Institute for Health and Care Excellence (NICE). *Guidance and advice list - Hepatitis C*; 2016. Available from: <https://www.nice.org.uk/guidance/published?type=ta&title=hepatitis%20C>.

67. Unverzagta S, Prondzinsky R, Peinemannc F. Single-center trials tend to provide larger treatment effects than multicenter trials: a systematic review. *J Clin Epidemiol* 2013;66:1271-80.

68. Bellomo R, Warrillow SJ, Reade MC. Why we should be wary of single-center trials. *Crit Care Med* 2009;37:3114-19.

69. Sinhaa IP, Sinhab SK. Single-center trials in neonatology: issues to consider. *Semin Fetal Neonatal Med* 2015 20:384-8.

70. Alexander PE, Bonner AJ, Agarwal A, Li SA, Hariharan At, Izhar Z, et al. Sensitivity subgroup analysis based on single-center vs. multi-center trial status when interpreting meta-analyses pooled estimates: the logical way forward. *J Clin Epidemiol* 2016;74:80-92.

71. Barbour V, Burch D, Godlee F, Heneghan C, Lehman R, Perera R, et al. Characterisation of trials where marketing purposes have been influential in study design: a descriptive study. *Trials* 2016;17:31.

72. Elston J, Taylor RS. Use of surrogate outcomes in cost-effectiveness models: a review of United Kingdom health technology assessment reports. *Int J Technol Assess Health Care* 2009;25:6-13.

73. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8:431-40.

74. Ellenberg SS, Hamilton JM. Surrogate endpoints in clinical trials: cancer. *Stat Med* 1989;8:405-13.

75. Herson J. The use of surrogate endpoints in clinical trials (an introduction to a series of four papers). *Stat Med* 1989;8:403-04.

76. Freedman LS, Graubard BL, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992;11:167-78.

77. de Gruttola V, Fleming T, Lin DY, Coombs R. Perspective: validating surrogate markers - are we being naive? *J Infect Dis* 1997;175:237-46.

78. Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliairy endnpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* 1994;13:955-68.

79. Fleming TR, De Mets DL. Surrogate end points in clinical trials: are we being misled? *Ann Int Med* 1996;125:605-13

80. Holloway RG, Dick AW. Clinical trial end points. On the road to nowhere? *Neurology* 2002;58:679-86.

81. Schievink B, Lambers Heerspink H, Leufkens H, De Zeeuw D, Hoekman J. The use of surrogate endpoints in regulating medicines for cardio-renal disease: opinions of stakeholders. *PLoS One* 2014;9:e108722.

82. U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER), U.S. Department of Health and Human Services. Food and Drug Administration. Center for Biologics Evaluation and Research (CBER). *Clinical trial endpoints for the approval of non-small cell lung cancer drugs and biologics: guidance for Iidustry*. Food and Drug Administration,; 2015. Available from: <http://www.fda.gov/downloads/Drugs/.../Guidances/UCM259421.pdf>.

83. European Medicines Agency. *CHMP assessment report Xalkori (crizotinib)*: European Medicines Agency. Committee for Medicinal Products for Human Use; 2012. Available from: <http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Summary_for_the_public/human/002489/WC500134762.pdf>

84. European Medicines Agency. *CHMP assessment report Unituxin (dinutuximab)*: European Medicines Agency. Committee for Medicinal Products for Human Use; 2015. Available from: <http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002800/WC500192794.pdf>

85. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users’ guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. *JAMA* 1999;282:771–8.

86. Lassere MN, Johnso KR, Boers M, Tugwell P, Brooks P, Simon L, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *J Rheumatol* 2007;34:607–15.

87. Berger VW. Does the Prentice criterion validate surrogate endpoints? *Stat Med* 2004;23:1571-78.

88. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998;54:1014-29.

89. Baker SG, Kramer BS. Surrogate endpoint analysis: an exercise in extrapolation. *J Natl Cancer Inst* 2013;105:316-20.

90. Zee J, Xie SX. Assessing treatment effects with surrogate survival outcomes using an internal validation subsample. *Clin Trial* 2015;12:333-41.

91. Ciani O, Davis S, Tappenden P, Garside R, Stein K, Cantrell A, et al. Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. *Int J Technol Assess Health Care* 2014;30:312-24.

92. Prasad V, Kim C, Burotto M, Vandross A. The strength of association between surrogate end points and survival in oncology: a systematic review of trial-level meta-analyses. *JAMA Intern Med* 2015;175:1389-98.

93. Ciani O, Buyse M, Garside R, Peters J, Saad ED, Stein K, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. *J Clin Epidemiol* 2015;68:833-42.

94. Blumenthal G, Karuri S, Zhang H, Zhang L, Khozin S, Kazandjian D, et al. Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non-small-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. *J Clin Oncol* 2015;33:1008-1.

95. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. *BMJ* 2013;346:f457.

96. Ciani O, Taylor RS. Surrogate, friend or foe? The need for case studies of the use of surrogate outcomes in cost-effectiveness analyses. *Health Econ* 2013;22:251-2.

97. Bujkiewicz S, Thompson JR, Sutton AJ, Cooper NJ, Harrison MJ, Symmons DP, et al. Use of Bayesian multivariate meta-analysis to estimate the HAQ for mapping onto the EQ-5D questionnaire in rheumatoid arthritis. *Value Health* 2014;17:109-15.

98. Bell H, Wailoo A, Hernandez M, Grieve R, Faria R, Gibson L, et al. *The use of real world data for the estimation of treatment effects in NICE decision making* London: National Institute for Health and Care Excellence (NICE). Decision Support Unit; 2016. Available from: <http://www.nicedsu.org.uk/RWD%20DSU%20REPORT%20Updated%20DECEMBER%202016.pdf>

99. Cunanan K, Gonen M, Shen R, Hyman D, Riely G, Begg C, et al. Basket trials in oncology: a trade-off between complexity and efficiency. *J Clin Oncol* 2017;35:271-73.

100. Renfro L, Sargent D. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Ann Oncol* 2016;Oct 11. pii: mdw413.

101. Simon R. Genomic alteration-driven clinical trial designs in oncology. *Ann Int Med* 2016;165:270-8.

102. *Accelerated Access Review: Final Report Review of innovative medicines and medical technologies*; 2016. Available from: <https://www.gov.uk/government/publications/accelerated-access-review-final-report>

103. National Institute for Health and Care Excellence (NICE). *TA395: Ceritinib for previously treated anaplastic lymphoma kinase positive non-small-cell lung cancer*; 2016 24th November. Available from: <https://www.nice.org.uk/Guidance/TA395>