ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

ISPOR Report

Methods for Evaluation of Surrogate Endpoints for Health Technology Assessment Decision Making: A Good Practices Report of an ISPOR Task Force

Sylwia Bujkiewicz, PhD,* Oriana Ciani, PhD,* Bart Heeg, PhD,* Dawn Lee, MMath, MSc, Jeanette M. Kusel, MSci, MSc, Kristian Thorlund, PhD, Petros Pechlivanoglou, PhD, Stephen Stefani, MD, Wanrudee Isaranuwatchai, PhD, Marc Buyse, PhD, Mario Ouwens, PhD

ABSTRACT

Surrogate endpoints are frequently used as primary outcomes in clinical trials. This is appropriate when they are validated for their ability to predict clinical benefit measured on patient-relevant target outcome(s). Such validation is often lacking, thus increasing uncertainty in the decision-making process of regulatory bodies, health technology assessment agencies and payers.
This ISPOR Task Force Report provides recommendations on best practices for surrogate endpoint evaluation for health technology assessment decision making. It covers methods that address the 3 levels of evidence for surrogate endpoint validation described in several methodological guidelines: (1) association between treatment effects on the surrogate and the target outcome, (2) association between the surrogate and the target outcome, and (3) biological plausibility. Statistical methods for surrogate endpoint evaluation include meta-analytic approaches using individual participant data or aggregate data. Multivariate meta-analytic models are recommended because they account for the within-study correlation and estimation errors. Issues with limited data and generalizability might be addressed through Bayesian approaches for information sharing from different treatments, treatment classes or indications. Real-world data can complement randomized controlled trial data, especially in rare diseases, but require careful consideration of underlying bias. For plausibility of health economic modeling, the surrogacy analysis and the health economic model should be aligned. The modeled time course of surrogate and target outcomes per treatment arm, as well as the modeled relative effects, should be reported to assess plausibility. Parameter and structural uncertainty in surrogate relationships can be explored through scenario analyses, probabilistic sensitivity analyses, value of information analyses, and threshold analysis techniques.

Keywords: outcomes research, surrogate endpoints, validation.

Cite this article as: Bujkiewicz S, Ciani O, Heeg B, et al. Methods for evaluation of surrogate endpoints for health technology assessment decision making: a Good Practices Report of an ISPOR Task Force. *Value Health.* 2026; 29(5):711–724.
doi.org/10.1016/j.jval.2026.01.020

Highlights

- This ISPOR Task Force report addresses how surrogate endpoints can statistically be evaluated and used in health technology assessment decision making. It provides integrated recommendations spanning both surrogate endpoint validation and health economic modeling, considering 3 levels of evidence (biological plausibility, individual-level, and trial-level surrogacy).
- The ISPOR Task Force recommends multivariate meta-analytic methods that account for within-study correlations and estimation errors when evaluating surrogate endpoints. Bayesian approaches can strengthen inferences by sharing information across treatments, classes, or indications when data are limited or generalizability is uncertain.
- The ISPOR task force recommends transparent, evidence-based use of surrogate endpoints in health technology assessment by aligning levels of evidence for surrogate endpoint validation with model-based cost-effectiveness—if this is expected—to enhance consistency and credibility in healthcare decision making.

Introduction

The use of surrogate endpoints to inform primary outcomes in clinical trials that support the regulatory approval of pharmaceutical and medical products is becoming increasingly common, in part due to the proliferation of expedited review programs. Approximately 60% of new drugs and biologics approved by the US Food and Drug administration (FDA) in the last 2 decades are based on surrogate endpoints.[1]

A surrogate endpoint does not measure the clinical benefit of primary interest in and of itself; instead, it is expected to predict clinical benefit or harm based on epidemiologic, therapeutic, pathophysiologic, statistical, or other scientific evidence.[2] Surrogate endpoints may encompass biomarkers, such as imaging findings or laboratory measurements, but also intermediate outcomes, such as exercise tolerance or time to disease progression, which can reduce both sample size and duration of clinical trials.

Validated surrogate endpoints (that have demonstrated predictive ability of clinical benefit) can serve as the basis for standard regulatory approval. Additionally, surrogate endpoints that are reasonably likely to predict clinical benefit may be used for

*Sylwia Bujkiewicz, Oriana Ciani, and Bart Heeg are joint first authors.

accelerated approval, pending postmarketing confirmatory trials to verify the anticipated claims.[3] However, recent studies found that most surrogate endpoints used to support FDA approval of drugs lacked high-level evidence of validation.[4-6]

The debate on the use of surrogate endpoints has traditionally evolved around regulatory considerations by bodies such as the FDA or the European Medicines Agency (EMA). However, trial findings are also essential to inform clinical practice and policy, and the potential use of surrogate endpoints extends beyond the regulatory setting to patients, clinicians, health technology assessment (HTA) organizations, payers, and other stakeholders.

Methodological recommendations for using surrogate endpoints vary considerably across HTA agencies (see Appendix A in Supplemental Materials drafted based on Grigore et al. and a recent white paper by international HTA agencies on surrogacy[7,8]). HTA organizations consider multiple statistical factors when it comes to evaluating surrogate endpoints, such as the strength of the association and the uncertainty around the association. Several HTA bodies apply the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework to assess certainty of evidence. According to GRADE, indirectness arises when a mismatch occurs between target population, intervention, comparators, and outcomes (PICO) elements and PICOs of the studies constituting the best available evidence; hence, surrogate endpoints typically trigger rating down the certainty of evidence for indirectness.[9]

The Institute for Quality and Efficiency in Health Care (IQWiG) in Germany is currently the only HTA agency that indicates a minimum numeric level for the (95% CI lower bound of the) correlation for determining whether an intermediate endpoint is a validated surrogate.[10] The National Institute for Health and Care and Excellence (NICE)'s Decision Support Unit Technical Support Document (TSD) 20 states that criteria for correlation may be unnecessary and that the focus should instead be on predictions and corresponding uncertainty.[11]

As such, HTA bodies differ in their views on validity of surrogate endpoints, causing overall lack of clarity for those developing and evaluating HTA reports.[12] Challenging situations occur when the technology under evaluation is the first of its kind (eg, new treatment class) or when multiple randomized control trials (RCTs) that have assessed both the surrogate and target outcome in the same drug class and indication are not available. Moreover, for HTA bodies that rely on cost-effectiveness models to establish the value of health technologies, guidance is needed on how to reflect in health economic analyses the uncertainty that results from not having sufficient information on the target outcome in the pivotal trial.

In 2024, the ISPOR Surrogate Endpoint Statistical Evaluation Task Force, a multistakeholder team of statisticians, health economists, regulators, decision makers, academics, and industry representatives convened to address these issues and to quantitatively characterize the relationship between surrogate endpoints and target outcomes.[13] This report provides guidance on the use of different surrogate evaluation methods in HTA decision making and on how surrogate endpoints and surrogate endpoint evaluations can be incorporated into cost-effectiveness models to predict target outcomes and corresponding uncertainty.

Levels of Evidence for Surrogate Endpoint Validity

Although there is considerable variability between the positions taken by different agencies, most organizations describe the evidence for surrogacy based on 3 levels outlined in Table 1. We use the illustrative example of low-density lipoprotein cholesterol as a surrogate for all-cause mortality.

We will discuss the 3 levels of evidence, starting from the necessary but not sufficient level 3 (biological plausibility) and then levels 1 and 2 (statistical validation). The evidence to inform validation should be gathered through systematic and documented approaches, as well as being specific to the population of interest and to the intervention type. We acknowledge that this may be difficult to achieve because of the limited number of studies available in some disease areas. In addition, the underlying studies should be robust according to well-known risk-of-bias assessment tools (eg, version 2 of the Cochrane risk-of-bias tool for randomized trials [RoB 2]).

Often due to the novelty of the technology, sufficiently mature evidence for the estimation of the treatment effects on both the surrogate endpoint and the target outcome may not yet be available, but this does not necessarily mean lack of surrogacy. Transferring evidence from a different population or intervention requires thorough justification and specific statistical approaches. Additionally, availability of aggregate or individual patient data (IPD), suitable real-world evidence, and the rarity of the condition should be considered to justify the selection of the most appropriate approach.

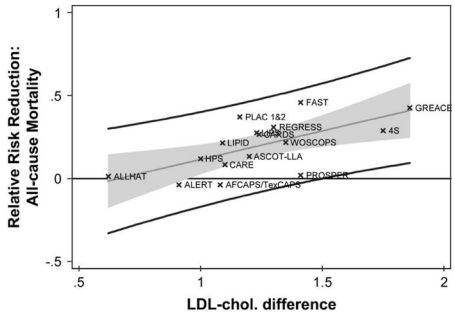
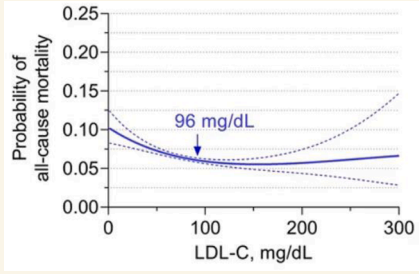
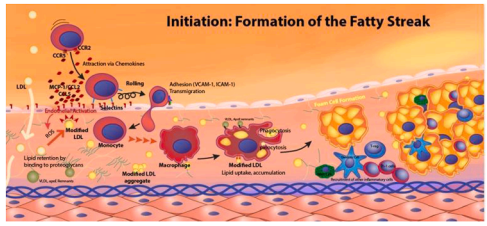
Biological Plausibility (Evidence Level 3)

A clear biological rationale is typically a prerequisite for the acceptability of a surrogate endpoint, especially in cases in which there is not yet statistical evidence that an effect on the surrogate endpoint translates into a treatment effect on the target outcome in the same drug class and/or population.[18] However, if robust data demonstrate no relationship between the surrogate endpoint and the target outcome, the assumed biological rationale becomes invalid. A biologically and clinically plausible surrogate relationship is one which is consistent, in terms of its existence, magnitude, and duration of effect, within the context of existing biomedical and epidemiological knowledge.

Plausibility should concern biological aspects, which are defined by disease processes, biological mechanisms and treatment mechanisms of initial and subsequent treatments and clinical aspects, mostly defined by human interaction with the biological process. For example, reduction of viral RNA in the blood below a certain threshold (ie, virological response), has been used as a biologically plausible surrogate endpoint in several infectious disease contexts to predict longer term outcomes known to be causally related to the disease.

One technique that could be useful in visualizing the biological reasoning for relationships between 1 or more potential surrogates and target outcomes is directed acyclic graphs (DAGs) (Fig. 1). In developing a DAG, it is important to consider the potential impact and underlying associations of confounding factors and/or treatment effect modifiers, such as baseline risk, time-varying characteristics, differences in mechanism of action, and anticipated treatment effects, on surrogate endpoint and target outcomes over time. Once assembled, the DAG and its depicted dependencies facilitate a more comprehensive understanding of the surrogate relationship, thereby informing future statistical analyses and supporting a structured consideration of how cost-effectiveness model inputs may influence both costs and quality-adjusted life-years (QALYs).[19,20] Although enhancing the transparency of the decision-analytic model and parameter selection, challenges such as oversimplification, limited data availability or quality, multiplicity of possible structures, may still hinder DAGs effective application.

Table 1. Levels of evidence for surrogate endpoint validity (Adapted From Ciani et al. 2017[14]).

Level of Evidence	Description	Example
Level 1: Statistical association between treatment effects on the surrogate endpoint and target outcome (trial-level association)	Several randomized controlled trials show that the effect of the intervention on the target outcome can be reliably predicted from the treatment effect on the surrogate endpoint.	 *
Level 2: Statistical association between surrogate endpoint and target outcome (individual-level association)	Observational studies (eg, cohort studies, case-control studies) or clinical trials show an association between the (change in the) surrogate endpoint and the target outcome at the individual level.	 †
Level 3: Biological plausibility	Pathophysiological studies and expert knowledge about the disease process indicate that the target outcome and surrogate endpoint are related and the treatment effect on the target outcome results from a treatment effect on the surrogate.	

LDL-C indicates low-density lipoprotein cholesterol.
Figures taken from.[15-17]
*From Johnson et al.[16] LDL-cholesterol change (mmol/L) was calculated by subtracting the 1-year LDL-cholesterol change in the control arm from the 1-year LDL-cholesterol change in the statin arm.
†From Li et al.[15] The primary exposure of interest was the first measurement of fasting lipid levels after hospitalization […] LDL-C was calculated by the Friedewald equation when triglyceride levels were below 400 mg/dL or otherwise measured directly.

Statistical Methods for Surrogate Endpoint Evaluation (Evidence Levels 1 and 2)

When evaluating surrogate endpoints, we are primarily interested in assessing the strength of the surrogate relationship between treatment effects on the surrogate endpoint and the target outcome (level-1 evidence for surrogacy, see Table 1). Causal pathways affecting such relationship are depicted in Appendix B in Supplemental Materials.

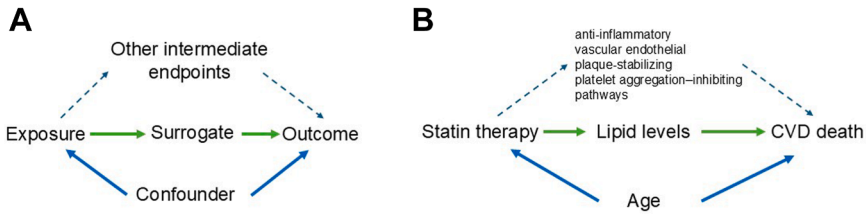
Statistical methods for evaluating surrogate endpoints as good predictors of clinical benefit have been studied extensively in the literature.[21,22] Many of the early methodological developments focused on evaluating surrogate endpoints based on data from a single trial, which has significant limitations.[23-27] Meta-analytic approaches for modeling the association between treatment effects on the surrogate endpoint and the target outcome, based on efficacy data from multiple trials[11,28-32] are more appropriate when evaluating surrogate endpoints at level-1 evidence.

In this section, we discuss statistical approaches for evaluation of surrogate endpoints at the levels of evidence discussed in the section Levels of Evidence for Surrogate Endpoint Validity, highlighting data requirements and methodological and data limitations. Although we focus here on issues that are specific to evaluation of surrogate endpoints, we note that following good practice guidelines for meta-analysis is recommended as in any literature review. This includes the assessment of risk of bias in included studies[33,34] and the overall quality of the body of evidence.[35]

Statistical Methods for Different Levels of Evidence

We begin by discussing meta-analytic methods allowing for evaluation of surrogate endpoints using both level-1 and -2 evidence simultaneously when sufficient data are available and then methods for evaluation of surrogate endpoints at level 1 or level 2 alone. The advantages and limitations of these approaches are highlighted in Table 2 and further discussed in the section

Figure 1. (A) Illustration of a directed acyclic graph (DAG) and (B) a DAG of the effect of statin therapy on cardiovascular disease (CVD) death. The effect of statins is mediated through lipid-level modification and through other mechanisms, such as through their antiinflammatory, vascular endothelial, plaque-stabilizing, and platelet aggregation-inhibiting effects. Age is considered as confounder and other factors (eg, diabetes) can affect the outcome independently of the exposure. Example modified from Dijk et al.[20]



Further Methodological Considerations and Extensions and Appendix C in Supplemental Materials, with examples of applying these methods in Appendix D in Supplemental Materials.

Methods for joint analyses of levels 1 and 2 evidence for surrogacy

Meta-analytic approaches utilizing IPD from multiple trials enable estimation of 2 statistical correlations (or other measures of association): the correlation between the surrogate endpoint and the target outcome (individual-level association; supporting level 2 evidence for surrogacy) and the correlation between the treatment effects on the surrogate endpoint and the target outcome (trial-level association; supporting level 1 evidence).[29]

Ideally, these 2 correlations (as well as the corresponding lower bounds) should be close to 1 for a surrogate endpoint to be statistically acceptable. Level-1 evidence is arguably more important for regulatory and HTA purposes because it allows one to predict the treatment effect on the target outcome having observed the treatment effect on the surrogate endpoint. These methods originally developed for continuous outcomes[29] were extended to time-to-event outcomes[36] and other types.[30] They represent the ideal situation in which complete analysis can be achieved for both levels. However, they require mature IPD from all trials in the meta-analysis. Issues around limited availability of IPD and approaches to surrogate endpoint evaluation when IPD are available only from a subset of trials are discussed in Appendix C.1 in Supplemental Materials.

Methods for level-1 evidence

When IPD are not available, aggregate data can be used to evaluate surrogate endpoints at the trial-level alone. Analyses based on aggregate data are easier to facilitate because such data are generally available from the existing literature. However, careful consideration needs to be given to the associated uncertainty and how the within-study association is accounted for. Figure 2 depicts 3 groups of statistical approaches for level-1 evidence with the corresponding assumptions made by each method.

The first approach, unweighted regression, is unsuitable for surrogate endpoint evaluation because it does not account for the measurement errors. The second approach, conventional meta-regression inversely weighted according to the variance of the estimated effect on the target outcome[46] has been used in the literature for surrogate endpoint evaluation. However, the method does not account for estimation errors around the treatment effects on the surrogate endpoint and ignores the within-trial association.

Weighted regression with weights proportional to the number of events or observations of the target outcome also does not accurately represent the uncertainty for both outcomes. Although meta-regression methods can serve the purpose of crude exploratory analysis, their above limitations can lead to underestimation of uncertainty and biased estimates of the trial-level association.[37–39,47]

Recommended methods that account for the within-study correlation and estimation errors for treatment effects on both the surrogate endpoint and the target outcome, depicted as methods group 3 in Figure 2 (which also include IPD based methods), include the bivariate meta-analytic fixed (independent) effects model[28] and its extensions to random effects.[11,39] These methods are described in the NICE Decision Support Unit Technical Support Document 20 (TSD20), in which code is made available in an appendix.[11] Some examples of such analyses are discussed in Appendix D.1 in Supplemental Materials. Approaches to the estimation of the within-study correlation (which is often not reported), some of which are summarized in TSD20, have been discussed by many authors.[11,37,48,49]

Evaluating the predictive value of surrogate endpoint at level-1 evidence

The predictive value of a surrogate endpoint (ability to predict the treatment effect on the target outcome from the treatment effect on the surrogate endpoint) may be evaluated by conducting a cross-validation procedure.[11,28] In addition, for IPD meta-analysis, datasets available at aggregate level alone and, therefore, excluded from the main analysis, can be used as validation data sets.[50] The focus on prediction is also embodied by the "surrogate threshold effect," the minimum effect on the surrogate endpoint that predicts a clinically meaningful effect on the target outcome.[51]

From the perspective of HTA decision making, prediction of a treatment effect on the target outcome from the effect on the surrogate endpoint may be of particular importance as a predicted effect may directly inform a decision model. Such predictions can be obtained utilizing the same meta-analytic models as those used for evaluation of the surrogate relationship.

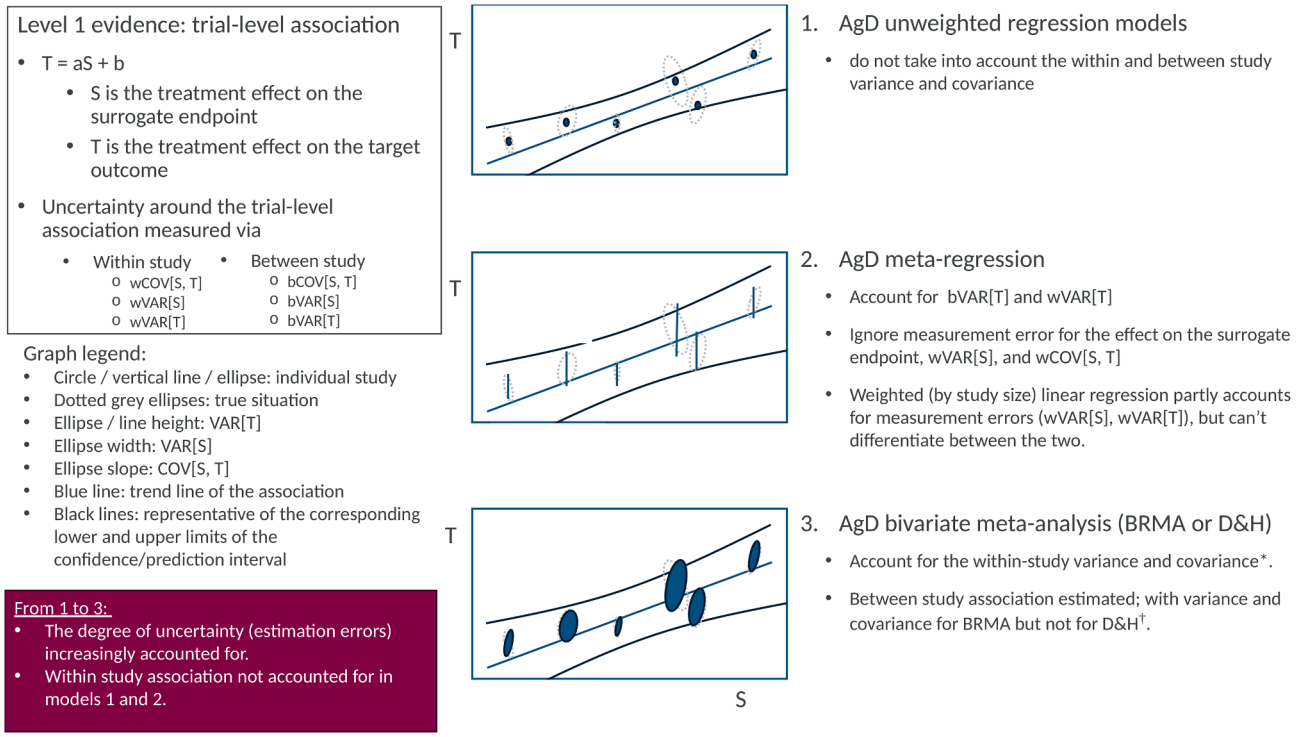
Methods for level-2 evidence for surrogate relationship

Although level-1 evidence is desirable when evaluating surrogate endpoints for HTA submissions, in some areas, such as in rare diseases or small subgroups of cancer patients positive for a targeted biomarker, RCT data may be scarce or not available at all. Data from a single RCT, a single-arm trial or an observational study (or a combination of those) may be used to estimate the correlation between the surrogate endpoint and the target

Table 2. Advantages and limitations of methods for evaluation of surrogate endpoints at level-1 and/or level-2 evidence.

Level	Method	Pros	Cons
Levels 1 and 2	IPD bivariate meta-analysis using regression-based mixed models[29,30,36]	• Models allow for estimation of both the individual-level and trial-level association. • Naturally account for measurement error. • Less strong distributional assumptions at the within-trial level (compared with methods for aggregate data). • Can be used to adjust for effect modifiers available in IPD datasets.	• Require IPD from all trials. • Assume bivariate normal distribution at the between-study level—sometimes a strong assumption as discussed below.
Level 1	Bivariate fixed (independent) effects meta-analysis with built-in regression.[28]	• Simpler methods using aggregate level data (no IPD required). • Account for measurement error for treatment effects on both the surrogate endpoint and the target outcome. • Account for the within-trial association. • Can be extended to adjust for covariates in a "meta-regression" style (for the random effects model a version in the product normal formulation can be used).	• Not possible to estimate individual-level association. • Difficulty of obtaining the within-study correlation between the treatment effects on the 2 outcomes (required to populate the model). - The within-study correlation is assumed known but rarely reported (IPD is the best source of data for estimation[28]). - Ignoring it may lead to biased results.[37,38] • Strong assumptions and transformations required for data other than continuous.
	Bivariate random effects meta-analysis.[11,39]		• Potential for increased uncertainty (due to fixed effects assumption) when the number of studies is small. • The assumption of random effects may not be valid for some data.[28] - It can be relaxed by replacing the normal distribution (at the between-study level) with a t-distribution.[39]
	Bayesian hierarchical meta-analysis models for borrowing information across treatment classes[40] or indications,[41] or network meta-analysis for borrowing information across treatment contrasts.[42]	• Ability to utilize a wider range of trials. • Improvement in precision around the estimates of the association quantifying the surrogate relationship in the class, treatment or indication of interest. • Improvement in precision around the predicted effect in a new study.	• Data needed from a relatively large number of trials and at least from 3 groups (classes or indications), with better performance with a larger number of groups. • Complex, particularly when not all units (classes or indications) are similar and partial exchangeability models need to be used to account for the different levels of similarity.[35,36]
	Meta-analytic methods for combining data from RCTs and non-randomized studies including bias adjustment.[43]	• Ability to utilize a wider range of trials. • Improvement in precision around the estimates of the association in the class/indication of interest. • Potential improvement in precision around the predicted effect in a new study.	• Risk of allocation bias and confounding.
Level 2	Regression models using RCT data for two correlated outcomes (models with correlated errors), including the treatment as a covariate; for example, individual-level part of the IPD meta-analytic regression-based mixed models.[29,30,36]	• Require fewer data (a single trial).	• Not possible to evaluate the association between the treatment effects. • Often data from observational studies are used, with higher risk of bias. • Often analysis restricted to estimating a correlation, which does not imply causation.
	Simple (IPD level) regression[44,45] of the target outcome with the surrogate endpoint as a covariate (or simple estimation of a correlation) based on data from an observational study and/or clinical trial.	• Require fewer data (a single study).	

Figure 2. Visual overview of statistical methods for level-1 evidence. AgD indicates aggregate data; BFMA, bivariate fixed/independent effects meta-analysis (model by Daniels and Hughes); BRMA, bivariate random effects meta-analysis; bCOV, between-study covariance; bVAR, between-study variance; IPD, individual patient data; wCOV, within-study covariance; wVAR, within-study variance.* Lack of random effects in BFMA (intercept, slope and variance of T conditional on S used as measures of association).



outcome. Here, IPD are preferable because aggregate data (used, for example, to estimate the association between the mean surrogate endpoint and the mean target outcome) may lack granularity to account for all relevant confounders or baseline risk factors or be prone to ecological bias.[52] Specific statistical methods will depend on the type of study the data are available from and will include different types of regression models or direct estimation of a correlation. They are discussed briefly in the bottom rows of Table 2.

Further Methodological Considerations and Extensions

Data availability, PICOs and methodological considerations

From a practical point of view, one of the greatest challenges for statistical evaluation of surrogate endpoints is the availability of relevant data on the effects of treatment on the surrogate endpoint and the target outcome (such as effect of trastuzumab on disease-free survival and overall survival [OS] in breast cancer—see an example in Appendix D.2 in Supplemental Materials).

The ideal situation is one in which multiple RCTs have been conducted to assess the effects of treatments of the same mechanism of action on the same surrogate endpoint and target outcome and in the same population. The number of trials must be sufficient for informative analyses to be possible.[53]

When evaluating surrogate endpoints in the HTA context, a review of existing evidence will reveal availability of data within the scope of a technology evaluation, typically defined by a set of PICOs. The scope for such review should be sufficient to capture all of the evidence required for evaluation of surrogate endpoints, which may need to be broader when data are limited within the main scope of the HTA. Some extended analyses discussed below may be attempted.

A desirable evidence base would include multiple RCTs with mature IPD available. But this is rarely feasible, and other approaches recommended above often provide acceptable solutions. The successful assessment of a surrogate relationship will depend on the existence of individual- and study-level associations but also the number and quality of available studies, number of events per treatment arm, variation of the treatment effect on the surrogate endpoint and the target outcome across studies and other methodological factors. Methodological assumptions, related limitations and some solutions to those limitations are discussed in Appendix C.2 in Supplemental Materials.

Generalizability and borrowing information from an extended evidence base

HTA organizations, including NICE, Pharmaceutical Benefits Advisory Committee in Australia, or the European HTA Coordination Group, advise against generalization of surrogate relationships to other treatments or populations.[42-44,54-56] However, Bayesian meta-analytic models are emerging for efficient use of diverse sources of data, including from different treatment classes,[40] indications,[41] and study designs,[43] when evaluating surrogate relationships while avoiding their direct generalization. These methods may particularly be useful in rare diseases or patient subgroups for which the evidence to evaluate surrogate endpoints in each setting alone is sparse.

Surrogate relationships may depend on the treatment class, specific treatment or even specific treatment contrast. For

example, the trial-level association may hold for placebo-controlled trials but not for trials with an active control. Data from many trials in a given setting would be required to investigate this. Bayesian bivariate network meta-analysis may provide a support here by allowing for evaluation of surrogate endpoints in different treatment contrasts while borrowing information across the contrasts.[42]

A Bayesian hierarchical model for modeling surrogate relationships within and across treatment classes allows for sharing information across therapies of different mechanism of action.[40] Both aforementioned methods, highlighted in the NICE methods guide,[54] often help with obtaining more precise estimates for the surrogate relationship. The hierarchical approach can also be utilized to share information across indications, which could substantially increase the evidence base due to sequential repurposing of drugs; for example, in different types of cancers.[41] Examples are included in Appendix D.3 in Supplemental Materials.

Real-world evidence

When data from RCTs for surrogate endpoint evaluation are limited, real-world data (RWD) can be useful to complement RCT data, albeit with appropriate consideration of sources of bias.[43] For example, in many rare diseases or for patients at the early stage of their disease, few RCTs boast sufficiently large sample sizes or long-enough follow-up time to observe a sufficient number of events.[57] When RWD from large nationwide cohorts are available for standard-of-care treatment options, with long-term follow-up, analysis of association between the surrogate endpoint and target outcome of interest may provide acceptable level-2 evidence.[58,59] However, such level-2 evidence should ideally be confirmed using high-quality data from subsequent RCTs. It is important to consider the target setting of the observational data, which will typically reflect routine clinical practice.

Use of RWD as level-1 evidence for surrogate endpoint evaluation is more questionable because there is no evidence that RWD can substitute for RCT data. When RWD are ideally available at IPD level, analysts should follow established frameworks for target trial emulation.[60-62] Strict adherence to the target trial protocol and comprehensive sensitivity analyses are required to ensure the transparency and robustness of analysis utilizing RWD at level-1 evidence because such methodological approaches and guidelines for analysts are still in development.[63] Careful use of RWD to supplement rather than replace RCT data may provide broader assessment of surrogacy for which RCT data are limited.[43] Further discussion of this topic is included in Appendix C.3 in Supplemental Materials.

Application of Methods and Data Availability

Although there are many examples of surrogate endpoint evaluation in clinical settings, such evaluations have been less common in HTA submissions, with manufacturers often stating data limitations.[64] Submissions referring to the literature on the validity of surrogate endpoints in clinical settings have often been criticized for the "extrapolation" of the surrogate relationship beyond the scope of a HTA (examples are given in Appendix E.1 in Supplemental Materials). Methods discussed in the section Further Methodological Considerations and Extensions may help alleviate some of these limitations. Figure 3 can guide the choice of methods depending on the type of data available, noting the need for bias assessment.[33,34,65] Additional examples of application of some of the methods discussed here that can be used to include surrogate endpoints in HTA decision modeling are discussed in Appendix E.2 in Supplemental Materials. Further methodological approaches are discussed in Appendix C.4 in Supplemental Materials.

Surrogate Endpoints in Health Economic Modeling

In this section, we explore the interplay between health economic modeling and surrogate endpoint evaluation. In doing so, we refer to HTA case studies to reflect on the use of level-1 and level-2 evidence in health economic models presented in Appendix F (Appendix Table F1 in Supplemental Materials).

Health Economic Modeling

In HTA, cost-effectiveness is often evaluated by estimating incremental costs and incremental QALYs across different interventions over a relevant time-horizon depending on duration of treatment impact.[66] Often an important driver of QALYs and costs is overall survival, which is also the most important target outcome. Survival is influenced by the disease process, treatment effects, baseline characteristics, setting, and treatment pathway.

Data on target outcomes, such as survival, are often limited or immature at the time of HTA for trials powered for surrogate or intermediate endpoints. To address this challenge, health economic models may incorporate links between surrogate endpoints and target outcomes to estimate cost-effectiveness.

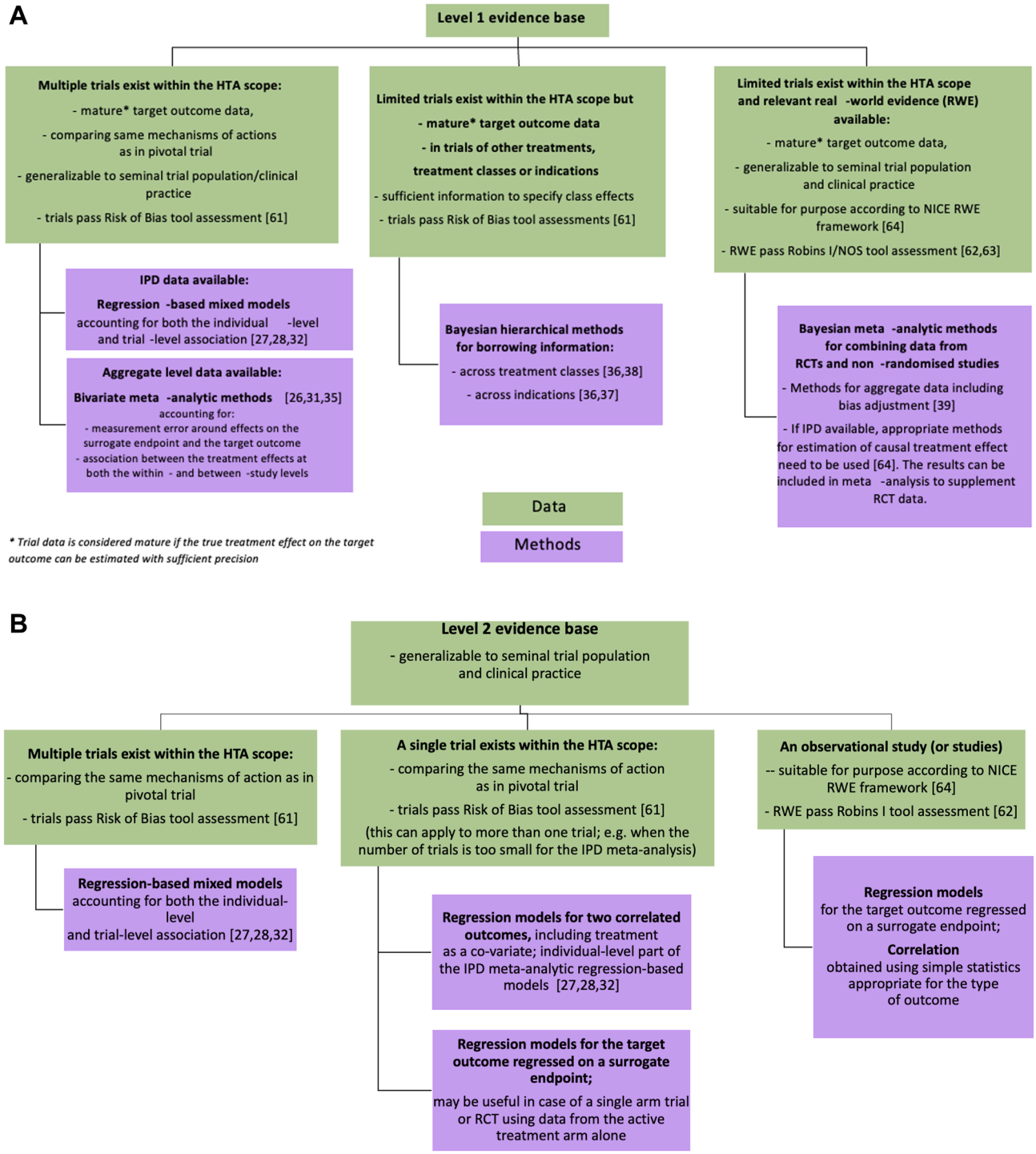
Evaluation of the 3 levels of surrogacy evidence, ie, biological plausibility, association between endpoints, and association between treatment effects on endpoints,[67] is therefore a necessary element for health economic modeling.

Biological Plausibility of Health Economic Models

As indicated before, DAGs can support the qualitative assessment of biological plausibility for clinical purposes. When designing health economic models, conceptual models may be developed, which typically are extended DAGs.[20,68] A quantitative approach to assess clinical/biological plausibility for overall survival based on health economic models is defined by Heeg et al[69] as predicted survival estimates that are considered reasonable based on prior expectations. These estimates are derived using methods that have been justified in advance.[68-70] The approach by Heeg et al provides a framework for developing and assessing the plausibility of health economic models and underlying data and associations, such as surrogacy associations.[69]

For biological plausibility assessment, it is important to consider both the within-trial period and extrapolation beyond the trial, ensuring that the model's structure, inputs and results remain plausible over time. For instance, reports should present the modeled estimates over time for all outcomes For time-to-event outcomes, this includes Kaplan-Meier curves, instantaneous hazards, and (time varying) hazard ratios with corresponding uncertainty over the modeled disease course. Reports should also include comparisons of modeled and observed survival data of the registrational trial. When available comparisons of external historical trial data with long-term follow-up and modeled data may be informative. With proper justification the external data may also inform survival prediction to increase plausibility of the extrapolations.

Figure 3. Diagram showing use of different methods (purple boxes) depending on data availability (green boxes) for (A) level-1 evidence and (B) level-2 evidence.



Approaches Used to Incorporate Level 1 and 2 Evidence in Health Economic Models

Level-1 and level-2 evidence may be used for model validation purposes or may be implemented directly in the model for prediction purposes. Below we will describe how level-1 and -2 evidence can be implemented in several health economic modeling approaches, including cohort state-transition models (StateTM), cohort partitioned survival analysis (PartSA), and patient-level simulations.[66,68,71-74]

State-transition models (StateTMs)

StateTMs are defined by health states and corresponding transition probabilities, which together determine survival.

Level-1 evidence is often not easily incorporated in StateTM because many surrogacy analyses are related to time from

randomization to event instead of time from health state entrance to event. An approach to developing fit for purpose level-1 evidence for StateTMs is to

- estimate relative treatment effects corresponding to each health state transition for each relevant trial,
- determine the relationship and strength of relationship (predictive value, correlation) between treatment effects on each health state transition and the treatment effect on the surrogate endpoint across all relevant trials,[75]
- and incorporate these surrogate relationships and their associated uncertainty into the health economic model.

This approach requires a large number of mature, relevant RCTs, which are rarely available.

The health state-transition probabilities in StateTMs are analogous to level-2 associations in surrogacy because both describe relations between end points. For the standard of care arm/anchor in slowly changing treatment landscapes, transitions from surrogate endpoint to survival can therefore usually be easily built in. These level-2 associations can be derived from systematic literature reviews or carefully selected databases adjusted to reflect the target setting.

Related to the treatment effect, StateTMs often assume no treatment effect on initial—and subsequent—health state mortality, assuming that any gains in time spent in the health state in which the treatment has demonstrated benefit translates to an OS gain. This assumption may not hold true. Delaying progression does not necessarily lead to an improvement in OS, particularly if effective subsequent lines of therapy are available. HTA bodies often request a scenario analysis where no OS benefit is assumed because this allows for the evaluation of the robustness of the model results under more conservative assumptions.[73]

Patient-level simulations

Similar to StateTMs, patient-level discrete event and state-transition-simulation models also naturally incorporate individual-level surrogacy (level-2 evidence), while accounting for other characteristics, such as baseline characteristics and other early outcomes. Related to level-1 evidence, often an assumption of homogeneity of treatment effects across simulated patients is made. This is, however, unrealistic when strong effect modifiers are present. The dependency of the treatment effect on effect modifiers needs to be accounted for. In general, this requires IPD for multiple trials underlying the surrogacy analyses.

Partitioned survival analysis (PartSA)

In PartSA, health state membership is determined by calculating the area between survival curves fitted to individual endpoints defined as time from randomization to an event. Endpoints other than OS are usually composite, such as progression-free survival, which combines progression and death events. Level-1 evidence can be implemented in a PartSA by using parametric distributions for the reference arm, to which the treatment effects on the surrogate outcome and predicted treatment effect on the target outcome are applied. In doing so, it is important to incorporate the entire surrogacy association, enabling scenario and sensitivity analyses that allow to easily account for the correlation between the surrogacy parameters.

Unlike StateTMs, level-2 evidence is not typically included in PartSA but may be implemented using methods such as landmark analyses for continuous/categorical surrogates[76,77] or via regression analyses linking composite time-to-event surrogate endpoints with death.[78,79] More information can be found in TSD21.[77] The validity of these methods depends on factors such as the timing of the landmark, the risk of selection bias, and the assumption that survival differences reflect the prognostic factor rather than underlying patient characteristics.

Considerations related to treatment effects in health economic models

It is important to note that models based on level-2 evidence to forecast a survival benefit are particularly susceptible to the surrogacy paradox,[80] whereby a demonstrated link between the early endpoint and survival does not translate into an actual survival gain.

Even when relying quantitatively on surrogate endpoints validated based on level-1 evidence, health economic models often require assumptions about the duration of treatment effects over a lifetime horizon. These assumptions are critical for model plausibility and can significantly affect cost-effectiveness conclusions and require justification and extensive testing in sensitivity analyses.

If No Level-1 Or Level-2 Evidence Exists

In some circumstances (particularly for orphan diseases), no level-1 and level-2 evidence may exist for modeling beyond the primary surrogate endpoint, such as in caplacizumab technology appraisal for treatment of acquired thrombotic thrombocytopenic purpura.[81] In these cases, it will be necessary to conduct economic analyses relying on biological plausibility to translate treatment effects on the surrogate endpoint to target outcome of interest. Heeg et al[69] provides some guidance on how to assess clinical and biological plausibility in the context of survival analysis. It is obviously crucial to involve clinicians and health authorities early in these cases to ensure they agree to the (conceptual) model, the assumptions and the underlying data. Structured expert elicitation may be helpful in defining plausible ranges for treatment effects based upon available information.[82] It is also important to try to collect additional evidence, which can be used directly in the model or as supportive evidence either for initial submission or to confirm results after a period of managed access. Finally, extensive sensitivity and scenario analyses are needed.

Uncertainty in Health Economic Modeling

The uncertainty inherent in the surrogate relationship needs to be explored and characterized as far as possible. Health economic models should account for both parameter uncertainty and structural uncertainty.[80,83-85]

- Parameter uncertainty arises from variability in parameter estimates, including surrogacy parameters.[79] The association between parameters and its uncertainty should be accounted for when exploring uncertainty.
- Structural uncertainty is uncertainty due to assumptions related to model structure and the generalizability of input data, affecting both point estimates and their associated uncertainty of, for instance, the applied level-2 and or -1 surrogacy associations.[80,83,85]

Although HTA bodies, such as NICE, recommend addressing structural uncertainties,[86] submissions frequently fail to do so

comprehensively, instead focusing on parameter uncertainty. As a result, true model uncertainty may be underestimated and/or potential bias in outcomes may be left unaddressed.

Probabilistic, 1-way sensitivity and scenario analysis are the most common methods used to assess the robustness of model outcomes.[80,85,86]

In the following sections, we outline how appropriate each method is for characterizing the impact of uncertainty in the surrogate relationship on the model results.

Probabilistic sensitivity analyses

Probabilistic sensitivity analysis (PSA) captures parameter uncertainty, including uncertainty in the surrogate relationship, while preserving the correlation structure between parameters.[84]

PSA can also be used to address structural uncertainty using methods, such as weighting and discrepancy approaches.[80,85]

- The weighting approach incorporates structural uncertainty by assigning probabilistic weights to different model structures or assumptions, reflecting their relative credibility based on expert opinion or empirical evidence. It is particularly useful when multiple modeling options exist.[83,85] For example, if 2 different risk equations linking the surrogate and target outcomes are available from separate sources, expert elicitation can be used to set expectations on the appropriateness of the risk equations defined by weights and uncertainty. These determine the proportion of PSA iterations in which each risk equation is applied.
- The discrepancy approach introduces discrepancy parameters summarizing both known and unknown structural uncertainty within the modeled relationships and their associated parameter uncertainty.[80] When using a level-1 surrogacy association to predict treatment effects on the target outcome, discrepancy parameters can adjust for generalizability issues affecting both the parameters describing the surrogacy relationship (eg, slope and intercept) and corresponding uncertainty.

For example, in the nivolumab case study, the authors applied a surrogacy association based on interferon trials, which may not fully apply to nivolumab, an immunotherapy (see Appendix Table F1 in Supplemental Materials). Structured expert elicitation can inform discrepancy parameters adjusting the surrogacy parameters, eg, slope and uncertainty, to set better expectations for nivolumab.

Incorporating structural uncertainty in PSA may enhance the quality of information presented to the decision maker.

After PSA, value of information analyses in terms of Expected Value of Perfect Parameter Information and Expected Value of Sample Information may be used to quantify the benefit of reducing uncertainty relating to the parameters for the surrogate relationship and to inform whether and which additional data collection efforts may be worthwhile to reduce this uncertainty.[87-89] For example, expected value of sample information can be used to determine whether the value of additional follow-up from a particular trial design justifies the additional costs of collection. To mitigate the risk of an incorrect decision, HTA bodies may consider conditional reimbursement and request extended pivotal trial follow-up or additional data collection, either before or after reimbursement. The emerging concept of living HTAs has also been proposed, particularly for therapies targeting chronic conditions, such as diabetes.[90]

One-way and multiway sensitivity analysis

Typically, surrogacy analyses are based on regression equations defined by multiple parameters. The estimates of these parameters are correlated. Considering these correlations in 1-way sensitivity analysis is important and can be done by for instance using Cholesky decompositions.

Multiway sensitivity analysis, which varies 2 or more inputs at the same time, can be used to explore how combinations of uncertainties affect model results including exploration of best- and worst-case scenarios.

Scenario analyses

Scenario analyses can be used to test different assumptions to explore the impact of structural uncertainty including best and worst-case scenarios. HTA authorities often require submission models to have the functionality to set the treatment effect on the target outcome to 0 along with other treatment effect waning scenarios.[86,91] This can easily be done for PartSA. In StateTMs, it is also possible to incorporate scenario analyses in which there are no differences in total life-years. By substantial reparameterization or rebuilding the model structure, one may be able to have equal survival probabilities over time. This may, however, lead to clinically implausible assumptions for some transitions.

Threshold analysis

Threshold analysis techniques, such as quantitative bias analyses (QBA)[92] or tipping point analysis, can be used to determine the values of the parameters for the surrogate relationship at which the intervention(s) is no longer cost-effective. Structured expert elicitation or literature can help determine whether such thresholds are clinically plausible. This provides valuable information for HTA decision makers to inform reimbursement decisions. For example, the minimum level of treatment effect on the surrogate endpoint needed to demonstrate cost-effectiveness at a specific incremental cost-effectiveness ratio threshold can be compared with the observed treatment impact.

Hurdles for Considering Level-1 Surrogacy Evidence in Health Economic Models

Few technology appraisals considered level-1 evidence in health economic models (see Appendix Table F1 in Supplemental Materials) because surrogate endpoint evaluation requires mature target outcome data from a considerable number of trials. These trials should ideally compare the same mechanisms of actions as the pivotal trial and exhibit sufficient variation in treatment effects. Such comprehensive trial data are rarely available.

To address this limitation, PICOs criteria informing surrogacy analyses may need to be broadened to include other indications or treatments with other mechanisms of action. In such cases, differences between drug classes, indications, and study designs (eg, RWE) should be accounted for. Bayesian hierarchical models for surrogate endpoint evaluation, combined with expert elicitation, can help account for these differences and quantify any potential bias.[40,41,43] This approach can enable the development of level-1 evidence that can inform or validate health economic models.

Whether HTA agencies will accept the broader surrogacy analyses likely varies from case to case. Therefore, we may still need to rely on extrapolation of immature trial data in PartSA or rely on level-2 associations in StateTM. Level-2 associations should be

Table 3. ISPOR Surrogate Endpoint Statistical Evaluation Good Practices Task Force recommendations.

Number	Recommendations for using statistical methods for surrogate endpoint evaluation
1	The surrogate endpoint should be fully defined and clarity given to what target outcome it is substituting and predicting for in relation to the population and treatment defined in the Population Intervention Comparator Outcome (PICO).
2	Justify the selection and comprehensiveness (preferably through a systematic review of the literature) and use of information sources for both the validation of surrogate endpoints, and the health economic modeling.
3	Account for (and justify) the generalizability of external sources of evidence used to inform the decision problem, consider the PICO and especially differences in population and treatments.
4	Evaluate the validity of surrogate endpoints by estimating the trial-level association between the treatment effects on surrogate endpoint and target outcome (across trials - level 1 evidence) and the individual-level association between the surrogate endpoint and the target outcome (level 2 evidence).
5	Utilize robust meta-analytic approaches, ideally with IPD across many relevant randomized controlled trials (RCTs), and, if IPD are not available, methods for aggregate level data. Methods should account for the measurement error for treatment effects on both surrogate endpoint and target outcome and for the within-study correlation.
6	Consider appropriate assumptions of specific methods developed in different contexts including, for example, different scales of outcomes (for example, an assumption of normality for log odds ratios (log ORs) may not always be appropriate).
7	Take into account uncertainty around the estimate of the treatment effects on the surrogate endpoint and the target outcome when modeling the trial-level association, and around the parameters of the statistical model when interpreting the results.
8	When data from RCTs within the scope of an HTA are limited, consider other relevant sources of evidence, including trials of different treatments or indications, utilizing more complex meta-analytic methods that allow for combining as well as differentiating between such heterogeneous data.
9	In circumstances of limited randomized trial data, consider supplementing the evidence base with real-world data (RWD), carefully considering the data quality and appropriate statistical methods for the analysis of such data to ensure minimization of selection bias and confounding.
Number	Recommendations for using surrogate endpoints in health economic modeling
1	The disease progression model used for the statistical evaluation of surrogacy (eg, directed acyclic graphs [DAGs]) should be aligned with the conceptual models informing the health economic model, with justification provided for any deviations. The alignment should consider prognostic factors, effect modification, proportional hazard violations and impact on other early outcomes, and impact of subsequent treatments. The health economic model should be built to reflect the disease pathway and the ways in which the intervention is expected to impact upon this pathway, including patient and caregiver outcomes. The conceptual model should not be designed around data availability for surrogate endpoints. Simplifications due to data availability should be justified.
2	The design of cost-effectiveness models should be assessed for clinical/biological plausibility. • Examine whether changes in modeled absolute and relative treatment effects are biologically/clinically plausible. Ensure transparent reporting of the model results for all relevant outcomes. Reports should present the modeled estimates over time for all surrogate and target outcomes. For time-to-event outcomes, this includes Kaplan-Meier curves, instantaneous hazards, and (time varying) hazard ratios with corresponding uncertainty over the modeled disease course. • For models relying on level 1 and/or level 2 evidence, the risk and underlying reasons why a treatment effect on the target outcome may or may not materialize (surrogacy paradox) should be qualitatively assessed and reported.
3	The effect of treatment on all transitions and the duration of those effects should be modeled using parameters that are modifiable within the health economic model. This especially concerns the relationship between the surrogate and the different transitions which should be included in the economic model. This should be performed in a way that does not overly harm the transparency and the run-time of the model.
4	Models relying on level-1 and/or level-2 evidence, rely on external data, and therefore, make assumptions about generalizability of the point estimates and uncertainty around them. Generalizability issues include the following: differences in patient populations, translation of data from treatments with different mechanisms of action, etc. When modeling, develop the simplest model that sufficiently reflects reality in terms of point estimates and uncertainty, and is functionally flexible enough to evaluate other plausible scenarios. If structural uncertainty is suspected, incorporate structural uncertainty analyses to assess generalizability. This could either take the form of scenario analysis or use of the discrepancy or weighting approaches in probabilistic sensitivity analyses (PSA).
5	In areas with limited long-term survival data in the pivotal trial and limited level-1 and -2 evidence in literature or real-world evidence databases—such as rare or orphan diseases—modeling beyond a surrogate primary endpoint should follow a totality-of-evidence approach, integrating biologically plausible rationale, available evidence, and expert opinion where necessary. Early engagement with clinicians and HTA bodies is essential to ensure alignment on the conceptual model and assumptions. Where feasible, efforts should be made to collect additional evidence to strengthen the model or support validation.
6	Parameter uncertainty around the level-1 and -2 surrogate relationship should be explored using PSA, OWSA in combination with multiway sensitivity analysis, and scenario analysis, accounting for correlation between surrogacy parameters.
7	Threshold analysis techniques, such as quantitative bias analyses (QBA) or tipping point analysis, can be used to determine the value of the surrogacy relationship parameters at which the intervention(s) is no longer cost-effective. Use threshold analysis to determine the treatment effect and treatment duration parameter settings that would result in a change in the reimbursement decision. For example, in an oncology state-transition model, one may vary the hazard ratio applied to the postprogression transition.

continued on next page

Table 3. Continued

Number	Recommendations for using statistical methods for surrogate endpoint evaluation
8	Managed access agreements or coverage with evidence development may help to manage the consequences of uncertainty in decision making based upon surrogate outcomes. Value of information (VOI) analyses can inform the decision to extend the trial follow-up or to perform post marketing research.
9	Models should be "future-proofed" designed to allow the efficient incorporation of new information on the surrogacy relationship or target outcome when new data becomes available.

based on systematic approaches, such as systematic literature reviews or systematic selected observational data sources. Structural uncertainty evaluation over these 3 different approaches is important as are sensitivity and scenario analyses.

It is also important to note that information on whether prior therapies of the same class have already demonstrated a survival benefit will influence the HTA authorities position to unconditionally recommend a novel therapy.

Recommendations from the ISPOR Surrogate Endpoint Statistical Evaluation Good Practices Task Force

Table 3 summarizes the recommendations from the ISPOR Task Force on Good practices for Surrogate Endpoint Statistical Evaluation, with the first part focusing on the statistical methods for evaluation of the validity of surrogate endpoints, and the second part on the use of surrogate endpoints in health economic modeling.

In addition to the recommendations presented above, the task force believe there is an ethical imperative for the IPD of all RCTs to be accessible for further analyses. A widespread lack of access to IPD from RCTs poses a major limitation for a complete validation of surrogate endpoints at both level-1 and -2 evidence. In addition, leveraging on the ongoing initiatives for HTA and regulatory alignment and for evidence requirements convergence,[93] we strongly encourage HTA agencies to agree on a common set of criteria for the assessment of surrogate endpoints. Currently, there is no consensus on statistical methods and criteria to validate a surrogate endpoint. An openly accessible repository maintained by a reputable and independent organization of high-quality surrogate endpoint validation studies, systematically organized by treatment, therapeutic class, indication, and substituted target outcome(s), would represent a valuable resource for researchers, clinicians, industry representatives, and decision makers.

Conclusion

The ISPOR Surrogate Endpoints Statistical Evaluation Good Practices Task Force sought to analyze current best practices and methods, together with adaptations for situations in which their practical application can be challenging, for the evaluation of surrogate endpoints in a HTA context with the overarching aim of improving the clinical effectiveness and cost-effectiveness assessment of new or existing drugs and other health technologies. The recommendations developed are intended for use by HTA agencies but also the wider community of HTA stakeholders including regulators, healthcare professionals, manufacturers, and clinical guidelines experts.

Please note that this report and the recommendations are based upon current methods, and these are continually evolving.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section. Dr Ouwens is an employee of AstraZeneca and has stock ownership and/or stock options or interests in the company. Dr Ciani is an editor for *Value in Health* and had no role in the peer-review process of this article.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2026.01.020.

Article and Author Information

Accepted for Publication: January 11, 2026

Published Online: March 27, 2026

doi: https://doi.org/10.1016/j.jval.2026.01.020

Author Affiliations: Biostatistics Research Group, Division of Public Health and Epidemiology, School of Medical Sciences, University of Leicester, Leicester, UK (Bujkiewicz); Center for Research on Health and Social Care Management, SDA Bocconi School of Management, Milan, Italy (Ciani); National Health Care Institute, Diemen, The Netherlands (Heeg); Peninsula Technology Assessment Group (PenTAG), University of Exeter, Exeter, England, UK (Lee); National Institute for Health and Care Excellence, London, England, UK (Kusel); Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada (Thorlund); Child Health Evaluative Sciences, The Hospital for Sick Children Research Institute, Toronto, Ontario, Canada (Pechlivanoglou); Grupo Oncoclinicas/Unimed Central RS, Porto Alegre, Brazil (Stefani); Health Intervention and Technology Assessment Program Foundation (HITAP), Thailand (Isaranuwatchai); Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, Ontario, Canada (Isaranuwatchai); CluePoints SA, Louvain-la-Neuve, Belgium (Buyse); IDDI, Louvain-la-Neuve, Belgium (Buyse); Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Hasselt, Belgium (Buyse); Real World Science and Analytics, Biopharmaceuticals, AstraZeneca, Gothenburg, Sweden (Ouwens).

Correspondence: Oriana Ciani, PhD, Center for Research on Health and Social Care management, SDA Bocconi School of Management, Milan, Italy. Email: oriana.ciani@unibocconi.it

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: The authors received no financial support for this research.

Acknowledgment: The authors express their sincere gratitude to Elizabeth Molsen-David for her invaluable guidance and support throughout the development of this taskforce report.

The task force's work was presented at ISPOR conferences in North America and Europe, where feedback was actively solicited. Furthermore, as part of the expert consensus development process, the report underwent 2 rounds of written review. Jona Lilienthal from the

Department of Medical Biometry, Institute for Quality and Efficiency in Health Care, Germany participated in the task force and contributed to this report. The authors thank the following for their written comments during the review rounds that improved the report: Saswata Paul Choudhury, Miranda Cooper, Maicon Falavigna, Lucia Fiestas-Navarrete (CDA), Zoe Garrett (NICE), Farah Husein (CDA), Jayeshkumar Kanani, Saskia Knies (ZIN), Yin Min Kyaw, Lincy Lal, Hwee-Lin Wee, Eric Morenz (CDA), Abiola Olaleye, Daniel Ollendorf (ICER), Mir-Masoud Pourrahmat, Paola Rivera Ramirez (IETSI), Rod S Taylor, Vassiki Sanogo, Yan Wang, and Tracy Westley. The authors also extend their appreciation to Leila Zakka for her contributions to the identification of relevant HTA reports.

Sylwia Bujkiewicz was funded by the Medical Research Council [MR/T025166/1] and supported by Leicester NIHR Biomedical Research Center [NIHR203327]. At the time of writing this report Bart Heeg was Vice President HEOR at Cytel, Weena 316, Rotterdam The Netherlands.

REFERENCES

1. Zhang AD, Puthumana J, Downing NS, Shah ND, Krumholz HM, Ross JS. Assessment of clinical trials supporting US Food and Drug Administration approval of novel therapeutic agents, 1995-2017. *JAMA Netw Open*. 2020;3(4):e203284.
2. FDA-NIH Biomarker Working Group, Biomarkers, EndpointS, and Other Tools (BEST). *Resource*. Silver Spring (MD): Food and Drug Administration (US); 2016. http://www.ncbi.nlm.nih.gov/books/NBK326791/. Accessed January 29, 2025.
3. Food and Drug Administration (FDA). Demonstrating substantial evidence of effectiveness for human drug and biological products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/demonstrating-substantial-evidence-effectiveness-human-drug-and-biological-products; 2019. Accessed March 3, 2026.
4. Gyawali B, Hey SP, Kesselheim AS. Evaluating the evidence behind the surrogate measures included in the FDA's table of surrogate endpoints as supporting approval of cancer drugs. *EClinicalmedicine*. 2020;21:100332.
5. Wallach JD, Yoon S, Doernberg H, et al. Associations between surrogate markers and clinical outcomes for nononcologic chronic disease treatments. *JAMA*. 2024;331(19):1646–1654.
6. Food and Drug Administration (FDA). Table of surrogate endpoints that were the basis of drug approval or licensure. https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure; 2024. Accessed March 3, 2026.
7. International Collaboration. *Surrogate Endpoints in Cost-Effectiveness Analysis for Use in Health Technology Assessment*. White Paper; 2024. https://www.cda-amc.ca/sites/default/files/MG%20Methods/surrogate-endpoints-report.pdf. Accessed March 3, 2026.
8. Grigore B, Ciani O, Dams F, et al. Surrogate endpoints in health technology assessment: an international review of methodological guidelines. *Pharmacoeconomics*. 2020;38(10):1055–1070.
9. Guyatt G, Iorio A, De Beer H, et al. Core GRADE 5: rating certainty of evidence-assessing indirectness. *BMJ*. 2025;389:e083865.
10. Institut für Qualität und Wirtschaftlichkeit im. Gesundheitswesen (IQWiG). Aussagekraft von Surrogatendpunkten in der Onkologie. (IQWiG-Berichte). Report No: Nr. https://www.iqwig.de/download/a10-05_rapid_report_surrogatendpunkte_in_der_onkologie.pdf; 2011. Accessed March 3, 2026.
11. Bujkiewicz S, Achana F, Papanikos T, Riley R, Abrams K. *Multivariate Meta-analysis of Summary Data for Combining Treatment Effects on Correlated Outcomes and Evaluating Surrogate Endpoints*. London: National Institute for Health and Care Excellence (NICE); 2019.
12. Ciani O, Grigore B, Blommestein H, et al. Validity of surrogate endpoints and their impact on coverage recommendations: a retrospective analysis across international health technology assessment agencies. *Med Decis Mak*. 2021;41(4):439–452.
13. Malone DC, Ramsey SD, Patrick DL, et al. Criteria and process for initiating and developing an ISPOR good practices task force report. *Value Health*. 2020;23(4):409–415.
14. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Time to review the role of surrogate end points in health policy: state of the art and the way forward. *Value Health*. 2017;20(3):487–495.
15. Li S, Zhang W, Liu H. Association between lipid levels and all-cause and cause-specific mortality in critically ill patients. *Sci Rep*. 2023;13(1):5109.
16. Johnson KR, Freemantle N, Anthony DM, Lassere MND. LDL-cholesterol differences predicted survival benefit in statin trials by the surrogate threshold effect (STE). *J Clin Epidemiol*. 2009;62(3):328–336.
17. Linton MF, Yancey PG, Davies SS, et al. *The role of lipids and lipoproteins in atherosclerosis*. South Dartmouth (MA): National Center for Biotechnology Information (NIH); 2000. http://www.ncbi.nlm.nih.gov/books/NBK343489/. Accessed March 13, 2025.
18. Gladwell D, Ciani O, Parnaby A, Palmer S. Surrogacy and the valuation of ATMPs: taking our place in the evidence generation/assessment continuum. *Pharmacoeconomics*. 2024;42(2):137–144.
19. Kaltenthaler E, Tappenden P, Paisley S, Squires H. *Identifying and Reviewing Evidence to Inform the Conceptualisation and Population of Cost-effectiveness Models*. London: National Institute for Health and Clinical Excellence (NICE); 2011.
20. Dijk SW, Korf M, Labrecque JA, et al. Directed acyclic graphs in decision-analytic modeling: bridging causal inference and effective model design in medical decision making. *Med Decis Mak*. 2025;45(3):223–231.
21. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat Med*. 2006;25(2):183–203.
22. Ensor H, Lee RJ, Sudlow C, Weir CJ. Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(5):859–879.
23. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989;8(4):431–440.
24. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605–613.
25. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998;54(3):1014.
26. Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol*. 2003;3(1):16.
27. Alonso A, Van Der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*. 2015;71(1):15–24.
28. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med*. 1997;16(17):1965–1982.
29. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000;1(1):49–67.
30. Burzykowski T, Molenberghs G, Buyse M, eds. *The Evaluation of Surrogate Endpoints*. New York, NY: Springer; 2005:408.
31. Alonso Abad A, Bigirumurame T, Burzykowski T, et al. *Applied Surrogate Endpoint Evaluation Methods With SAS and R*. New York, NY: CRC Press; 2017:373.
32. Xie W, Halabi S, Tierney JF, et al. A systematic review and recommendation for reporting of surrogate endpoint evaluation using meta-analyses. *JNCI Cancer Spectr*. 2019;3(1):pkz002.
33. Higgins JPT, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 1st ed. Chichester, UK: Wiley; 2019.
34. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
35. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. https://www.bmj.com/content/336/7650/924; 2008. Accessed July 24, 2025.
36. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *J R Stat Soc C*. 2001;50(4):405–422.
37. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *J R Stat Soc A*. 2009;172(4):789–811.
38. Papanikos T, Thompson JR, Abrams KR, Bujkiewicz S. Use of copula to model within-study association in bivariate meta-analysis of binomial data at the aggregate level: a Bayesian approach and application to surrogate endpoint evaluation. *Stat Med*. 2022;41(25):4961–4981.
39. Bujkiewicz S, Thompson JR, Spata E, Abrams KR. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. *Stat Methods Med Res*. 2017;26(5):2287–2318.
40. Papanikos T, Thompson JR, Abrams KR, et al. Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data. *Stat Med*. 2020;39(8):1103–1124.
41. Singh J, Anwer S, Palmer S, et al. Multi-indication evidence synthesis in oncology health technology assessment: meta-analysis methods and their application to a case study of bevacizumab. *Med Decis Mak*. 2025;45(1):17–33.
42. Bujkiewicz S, Jackson D, Thompson JR, et al. Bivariate network meta-analysis for surrogate endpoint evaluation. *Stat Med*. 2019;38(18):3322–3341.
43. Wheaton L, Papanikos A, Thomas A, Bujkiewicz S. Using Bayesian evidence synthesis methods to incorporate real-world evidence in surrogate endpoint evaluation. *Med Decis Mak*. 2023;43(5):539–552.
44. Agresti A. *Foundations of Linear and Generalized Linear Models*. Hoboken, New Jersey, NJ: John Wiley & Sons Inc; 2015.
45. Collett D. *Modeling Survival Data in Medical Research*. 4th ed. Boca Raton: Chapman & Hall/CRC; 2023. https://www.taylorfrancis.com/books/9781003282525. Accessed April 1, 2025.
46. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559–1573.
47. Vickers AD. A comparison of the performance of 6 surrogacy models, including weighted linear regression, meta-regression, and bivariate meta-analysis. *Value Health*. 2025;28(4):591–598.
48. Collier W, Haaland B, Inker L, Greene T. Handling missing within-study correlations in the evaluation of surrogate endpoints. *Stat Med*. 2023;42(26):4738–4762.
49. Wei Y, Higgins JP. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Stat Med*. 2013;32(7):1191–1205.
50. Buyse M, Molenberghs G, Paoletti X, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biom J*. 2016;58(1):104–132.

51. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat*. 2006;5(3):173–186.
52. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points—the challenge of statistical validation. *Nat Rev Clin Oncol*. 2010;7(6):309–317.
53. Dimier N, Todd S. An investigation into the two-stage meta-analytic copula modeling approach for evaluating time-to-event surrogate endpoints which comprise of one or more events of interest. *Pharm Stat*. 2017;16(5):322–333.
54. National Institute for Health and Care Excellence (NICE). *NICE Health Technology Evaluations: the Manual*. London: National Institute for Health and Care Excellence (NICE); 2025.
55. HTA CG. Guidance on outcomes for Joint Clinical Assessments. https://www.eunethta.eu/wp-content/uploads/2018/01/Endpoints-used-in-Relative-Effectiveness-Assessment-Surrogate-Endpoints_Amended-JA1-Guideline_Final-Nov-2015.pdf; 2015. Accessed March 3, 2026.
56. Australian Government Department of Health and Aged Care. *PBAC guidelines | Appendix 5 Translating comparative treatment effects of proposed surrogate measures to target clinical outcomes*. Australian Government Department of Health and Aged Care; 2025. https://pbac.pbs.gov.au/appendixes/appendix-5.html. Accessed February 24, 2025.
57. Thorlund K, Shephard C, Machado L, et al. Adapting Health Technology Assessment Agency standards for surrogate outcomes in early stage cancer trials: what needs to happen? *Expert Rev Pharmacoecon Outcomes Res*. 2024;24(3):331–342.
58. Antonini M, Mattar A, Bauk Richter FG, et al. Real-world evidence of neoadjuvant chemotherapy for breast cancer treatment in a Brazilian multicenter cohort: correlation of pathological complete response with overall survival. *Breast*. 2023;72:103577.
59. Simon TG, Roelstraete B, Hagström H, Loomba R, Ludvigsson JF. Progression of non-alcoholic fatty liver disease and long-term outcomes: a nationwide paired liver biopsy cohort study. *J Hepatol*. 2023;79(6):1366–1373.
60. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758–764.
61. Zuo H, Yu L, Campbell SM, Yamamoto SS, Yuan Y. The implementation of target trial emulation for causal inference: a scoping review. *J Clin Epidemiol*. 2023;162:29–37.
62. Hernán MA, Wang W, Leaf DE. Target trial emulation: a framework for causal inference from observational data. *JAMA*. 2022;328(24):2446.
63. Gomes M, Latimer N, Soares M, et al. Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *Pharmacoeconomics*. 2022;40(6):577–586.
64. Wheaton L, Bujkiewicz S. Use of surrogate endpoints in health technology assessment: a review of selected NICE technology appraisals in oncology. *Int J Technol Assess Health Care*. 2025;41(1):e11.
65. Farrah K, Young K, Tunis MC, Zhao L. Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. *Syst Rev*. 2019;8(1):280.
66. Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–1. *Value Health*. 2012;15(6):796–803.
67. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Use of surrogate end points in healthcare policy: a proposal for adoption of a validation framework. *Nat Rev Drug Discov*. 2016;15(7):516.
68. Kaltenthaler E, Tappenden P, Paisley S, Squires H. *NICE DSU Technical Support Document 13: identifying and Reviewing Evidence to Inform the Conceptualisation and Population of Cost-Effectiveness Models*. London: National Institute for Health and Care Excellence (NICE); 2011.
69. Heeg B, Lee D, Adam J, Postma M, Ouwens M. Defining biological and clinical plausibility: the DICSA framework for protocolized assessment in survival extrapolations across therapeutic areas. *Pharmacoeconomics*. 2025;43(7):793–803.
70. Roberts M, Russell LB, Paltiel AD, et al. Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–2. *Value Health s*. 2012;15(6):804–811.
71. Davis S, Stevenson M, Tappenden P, Wailoo A. NICE DSU Technical support document 15: cost-effectiveness modeling using patient-level simulation. Report by the Decision Support Unit. https://www.ncbi.nlm.nih.gov/books/NBK310370/pdf/Bookshelf_NBK310370.pdf; 2014. Accessed January 21, 2025.
72. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. *Med Decis Mak*. 2012;32(5):701–711.
73. National Institute for Health and Care Excellence (NICE), Decision Support Unit (DSU). *Technical Support Document 19. Partitioned Survival Analysis for Decision Modelling in Health Care: A Critical Review*. London: National Institute for Health and Care Excellence (NICE); 2018.
74. Siebert U, Alagoz O, Bayoumi AM, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–3. *Value Health*. 2012;15(6):812–820.
75. Jansen JP, Incerti D, Trikalinos TA. Multi-state network meta-analysis of progression and survival data. *Stat Med*. 2023;42(19):3371–3391.
76. National Institute for Health and Care Excellence (NICE). *Bortezomib for induction therapy in multiple myeloma before high-dose chemotherapy and autologous stem cell transplantation | Guidance*. National Institute for Health and Care Excellence (NICE); 2014. https://www.nice.org.uk/guidance/ta311/history. Accessed January 29, 2025.
77. Rutherford MJ, Lambert PC, Sweeting MJ, Pennington R, Crowther MJ, Abrams KR. *NICE DSU technical support document 21: flexible methods for survival analysis. Decision Support Unit*. University of Sheffield; 2020. https://www.sheffield.ac.uk/sites/default/files/2022-02/TSD21-Flex-Surv-TSD-21_Final_alt_text.pdf. Accessed January 29, 2025.
78. Dimopoulos M, Sonneveld P, Manier S, et al. Progression-free survival as a surrogate endpoint for overall survival in patients with relapsed or refractory multiple myeloma. *BMC Cancer*. 2024;24(1):541.
79. Félix J, Aragão F, Almeida JM, et al. Time-dependent endpoints as predictors of overall survival in multiple myeloma. *BMC Cancer*. 2013;13(1):122.
80. Strong M, Oakley JE, Chilcott J. Managing structural uncertainty in health economic decision models: a discrepancy approach. *J R Stat Soc C*. 2012;61(1):25–45.
81. National Institute for Health and Care Excellence (NICE). *Caplacizumab with plasma exchange and immunosuppression for treating acute acquired thrombotic thrombocytopenic purpura. Guidance*. National Institute for Health and Care Excellence (NICE); 2020. https://www.nice.org.uk/guidance/TA667/chapter/1-Recommendations. Accessed June 11, 2025.
82. University of Sheffield, et al. *The Decision Support Unit (DSU) is run by the University of Sheffield, providing advice and support to the National Institute for Health and Care Excellence (NICE)*. National Institute for Health and Care Excellence (NICE); 2025. http://www.nicedsu.org.uk. Accessed March 3, 2026.
83. Jackson CH, Sharples LD, Thompson SG. Structural and parameter uncertainty in Bayesian cost-effectiveness models. *J R Stat Soc C*. 2010;59(2):233–253.
84. Briggs AH, Weinstein MC, Fenwick EAL, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. *Med Decis Mak*. 2012;32(5):722–732.
85. Jackson CH, Bojke L, Thompson SG, Claxton K, Sharples LD. A framework for addressing structural uncertainty in decision models. *Med Decis Mak*. 2011;31(4):662–674.
86. National Institute for Health and Care Excellence (NICE). *Overview | NICE health technology evaluations: the manual | Guidance*. National Institute for Health and Care Excellence (NICE); 2022. https://www.nice.org.uk/process/pmg36. Accessed January 29, 2025.
87. Fenwick E, Steuten L, Knies S, et al. Value of information analysis for research decisions-an introduction: report 1 of the ISPOR value of information analysis emerging good practices task force. *Value Health*. 2020;23(2):139–150.
88. Rothery C, Strong M, Koffijberg HE, et al. Value of information analytical methods: report 2 of the ISPOR value of information analysis emerging good practices task force. *Value Health*. 2020;23(3):277–286.
89. Strong M, Oakley JE, Brennan A, Breeze P. Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Med Decis Mak*. 2015;35(5):570–583.
90. Daly MJ, Elvidge J, Chantler T, Dawoud D. A review of economic models submitted to NICE's technology appraisal programme, for treatments of T1DM & T2DM. *Front Pharmacol*. 2022;13:887298.
91. National Institute for Health and Care Excellence (NICE). *Surrogate endpoints in cost-effectiveness analysis for use in health technology assessment*. National Institute for Health and Care Excellence (NICE), CDA-AMC, PBAC, and ZIN; 2024. https://www.nice.org.uk/about/what-we-do/our-research-work/our-projects-and-partners#surrogate-endpoints-in-cost-effectiveness-analysis-for-use-in-health-technology-assessment. Accessed March 3, 2026.
92. Leahy TP, Duffield S, Kent S, et al. Application of quantitative bias analysis for unmeasured confounding in cost-effectiveness modeling. *J Comp Eff Res*. 2022;11(12):861–870.
93. Wang T, McAuslane N, Goettsch WG, Leufkens HGM, De Bruin ML. Regulatory, health technology assessment and company interactions: the current landscape and future ecosystem for drug development, review and reimbursement. *Int J Technol Assess Health Care*. 2023;39(1):e20.