Modelling uncertain heterogeneity for decision analytic models: an early exploration

Journal Title
XX(X):2–??
©The Author(s)
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Chengyang Gao

Abstract

Background: Health economic evaluations are essential for assessing the comparative costs and benefits of new health technologies. A critical aspect of these evaluations is accounting for heterogeneous treatment effects (HTE), which can greatly impact decision-making. Nevertheless, efficacy trials often lack sufficient statistical power and covariate coverage to characterize HTEs, leading many evaluations to assume constant effects.

Methods: We conduct a simulation study where the analysis of a randomised controlled trial informs decision modelling with a three-state state transition model. We compare five outcome modelling approaches under various outcome generating processes: unadjusted, adjusted, linear interaction, unrestricted spline, and monotonic spline models. The modelling approaches are assessed on their ability to predict population-level incremental net monetary benefits and their performance in probabilistic cost-effectiveness predictions in the target population.

Results: Flexible spline-based models with structured priors are better calibrated across a range of HTE scenarios and provide more appropriate representation of extrapolation uncertainty. In contrast, models assuming constant relative effect or linear interactions frequently produced overconfident predictions, especially when treatment effects declined in the extrapolated region. Point prediction of population-average INMB proved challenging under limited data. All models yielded similar level of predictive performance with large uncertainty intervals.

Conclusion: Modelling HTE, even under limited supporting evidence, can mitigate overconfidence in decision-making. Modellers should adopt flexible modelling approaches with careful considerations of regularisation and extrapolations. The lack of accuracy in point predictions also highlights the importance of communicating the uncertainty when the supporting evidence is limited.

Keywords

Introduction

Health economic evaluations systematically assess the comparative costs and benefits of emerging health technologies. Broadly, these evaluations employ decision-analytic models to synthesize comparative economic and clinical evidence for a new intervention against existing alternatives[1]. The rapid pace of innovation in health technologies has only increased the importance of these evaluations in guiding health resource allocation.

Variation in outcomes across individuals — broadly referred to as *heterogeneity* — must be accounted for to ensure fair and accurate evaluations. Heterogeneity arises in various forms, including differences in baseline risk and treatment effects (*clinical heterogeneity*), individual preferences, and patterns of resource use[2]. Ignoring these differences can introduce 'heterogeneity bias'[3–6]. Recognising this, HTA bodies such as NICE advocate for subgroup-specific cost-effectiveness analyses when heterogeneous treatment effects (HTEs) are justified[7]. Complementing this, machine learning has recently been used to flexibly examine heterogeneous baseline risks and treatment effects using individual participant data (IPD) from trials[8–10]. However, these analyses usually operate within the trial sample, which may not reflect the broader target population.

Despite growing awareness, many evaluations continue to assume a homogeneous cohort and focus on average treatment effects (ATE), especially when evidence for HTEs is limited. This is partly pragmatic: single trials often lack the statistical power to detect treatment-covariate interactions, meaning that IPD meta-regression is the current gold standard for identifying HTEs.[11–13] However, this may not be feasible for newly approved interventions, where the supporting evidence comes from efficacy trials. These efficacy trials often involve populations that are more selective and homogeneous than the wider population[14]. In such cases, analysts and regulators often default to a conservative approach, avoiding modelling HTEs due to insufficient evidence. Regulatory caution reinforces this preference for population-average decisions when causal evidence is lacking[15].

Corresponding author:

Email: chengyang.gao.15@ucl.ac.uk

Still, the rationale for considering HTEs extends beyond stratified decisions. The primary task of health economic modelling is to estimate population-level costs and benefits. Doing so requires generalising trial findings to the broader, heterogeneous population. Even with limited trial data, modelling plausible HTEs can support realistic extrapolation. In contrast, relying on sample average treatment effects (SATEs) assumes constancy across the population and may misrepresent the benefits of the intervention.

In our view, there are important nuances in whether to model HTEs with limited evidence, particularly in trial-based economic evaluations. We advocate for a predictive modelling perspective, focusing on probabilistic predictions of average treatment benefits across the target population. From this perspective, modelling HTEs and the associated uncertainty, are crucial for producing predictions that guide decision-making.

Methodologically, modelling and extrapolating uncertain HTEs presents substantial challenges. Often, it requires modelling the outcome and treatment effect surfaces from the trial sample. Without knowledge of their true shapes, data-driven approaches can yield misleading predictions when naively extrapolated beyond the observed data. This issue is especially pronounced in flexible outcome models[16]. Rather than neglecting uncertain HTEs,we argue for modelling approaches that constrain extrapolation behaviour. This can be achieved by combining a flexible outcome model with structured priors that reflect expert knowledge, using a fully Bayesian framework. The resulting uncertainty can then be quantified using Value of information (VoI)[17,18] analysis to prioritise research and design trials.

Unlike previous work that quantifies heterogeneity bias under 'known' subgroup effects[5], our case study investigates cost-effectiveness from different outcome modelling methods with uncertain HTEs. We consider a three-state state transition model, where transition probabilities are determined by outcome and treatment effect surfaces for a binary outcome. Our goal is to explore the value of modelling HTEs when evidence is limited. Using a simulation study, we compare the predictive performance of different modelling approaches across diverse outcome surfaces, using both decision-relevant and predictive metrics.

We begin by introducing the decision modelling framework and discussing how HTEs can be incorporated with limited supporting evidence. Section 3 explores modelling approaches for extrapolation. Section 4 details the simulation setup before the results and discussion are presented. Through this case study, we aim to contribute to ongoing discussions on modelling HTEs for decision-analytic modelling.

Decision analytic models and heterogeneity considerations

Decision analytic models (DAMs) provide a framework to synthesize clinical and economic evidence. They integrate data on disease progression, treatment effects and costs to estimate the population-average life-time for different interventions, even when only short-term data are available. As decision modelling aims to estimate population-average outcomes, clinical heterogeneity is often ignored, especially when it is uncertain. While there is an increasing recognition that accounting for heterogeneity is important for stratified decision-making, we consider the importance of considering heterogeneity even when targeting population-average outcomes.

We begin by providing a general introduction to DAMs before focusing on state transition models (STMs), a common type of DAM. Subsequently, we discuss the key considerations of heterogeneity in health economic modelling.

Overview of decision modelling

DAMs systematically evaluate the lifetime clinical benefits and costs associated with healthcare interventions[19]. Clinical benefit is often measured using quality-adjusted life-years (QALYs), a metric anchored at 0 for death and 1 for perfect health, while costs can be those incurred by the healthcare system and/or wider system costs[20]. To facilitate economic decision-making, healthcare payers conventionally adopt a willingness-to-pay (WTP) threshold (or a range), denoted by k, representing the maximum amount a decision-maker is prepared to pay for a unit gain in health benefits (e.g., per additional QALY). The net monetary benefit (NMB) of an intervention j is defined as

$$\mathrm{NMB}_j(\boldsymbol{\theta}) = k \cdot e_j(\boldsymbol{\theta}) - c_j(\boldsymbol{\theta}), \tag{1}$$

where $e_j(\boldsymbol{\theta})$ and $c_j(\boldsymbol{\theta})$ denote the cumulative QALYs and costs associated with intervention j, given parameter values $\boldsymbol{\theta}$.

Parameters in DAMs combine to determine intervention-specific net monetary benefit, enabling the calculation of incremental net monetary benefits (INMB)[21] between competing interventions. For simplicity, but without loss of generality, INMB of an intervention j relative to a reference treatment (e.g., $j = 1$) is defined as

$$\mathrm{INMB}_j(\boldsymbol{\theta}) = \mathrm{NMB}_j(\boldsymbol{\theta}) - \mathrm{NMB}_1(\boldsymbol{\theta}), \tag{2}$$

with a positive INMB indicating that intervention j is cost-effective at the specified threshold level.

Typically, these parameters are synthesized from different data sources with varying precision. Thus, a critical element of DAMs is the characterization of parameter uncertainty[22], often achieved using probabilistic sensitivity analysis (PSA)[23–25]. It has been argued that the Bayesian framework provides a natural approach to PSA[26]. This is because the posterior distribution for the parameters automatically captures parameter uncertainty and synthesizes the available evidence. The parameter uncertainty is then propagated through the decision model, yielding posterior distributions of costs and health benefits. We therefore adopt a Bayesian perspective for the remainder of this manuscript.

Cohort STMs are a common DAM and model the patient population as a single homogeneous cohort. This framework can be extended to multi-cohort STMs, which evaluate interventions across several cohorts with distinct characteristics while assuming homogeneity within each cohort. Individual-level STMs allow for increased heterogeneity based on individual characteristics while discrete event simulation (DES) incorporate individual interactions with the healthcare system[27,28]. Since the interaction between individuals and environment is not a main concern, we focus on STMs.

Heterogeneity considerations in decision modelling

Accounting for HTEs is relevant in all decision modelling, but has primarily been discussed in the context of stratified decision-making. We will introduce this theory and discuss the evidence required for this analysis. We then present a different perspective—modelling HTEs for generalising trial results to the target population. Finally, we argue that even when targeting population-average cost-effectiveness, modelling uncertain HTEs can still be valuable.

The incorporation of heterogeneity into DAMs has been extensively explored in the HTA literature[2,5,29,30]. These discussions usually presume that HTEs can be established based on current evidence. Thus, incorporating HTEs leads to stratified decision rules that achieve higher net benefits. Such policy stratification demands strong evidence—often network meta-regression[11–13]. Without such evidence, HTE modelling is usually deemed too uncertain or unreliable to support stratified decision-making. Trial-based decision modelling, specifically, is under-powered for HTE analysis and is often ignored.

But we argue that HTEs should be considered even when they cannot directly inform stratified decisions. The focus on stratified decisions has linked HTE modelling to a demanding decision problem — optimising decisions across subpopulations — which requires strong evidence within each subgroup. However, when decisions are considered at the overall population level, modelling HTEs are key to generalise trial findings.

The importance of generalisation for estimating the population-average cost-effectiveness is widely recognized, known as 'transportability' or 'external validity'[7,31–33]. In this context, ignoring HTEs altogether, even under limited supporting evidence, may lead to poorer decisions. Doing so effectively assumes the SATE equals the population average treatment effects (PATE). As clinical trial populations tend to be highly selective, the assumption can unrealistic.

Conversely, when guided by substantive knowledge, modelling HTEs can accurately represent uncertainty around treatment benefits in the target population. Propagating this uncertainty through the decision model, allows for a more accurate characterisation of decision uncertainty. Nevertheless, modelling HTE to generalise the trial data requires care, which we examine in the next section.

Modelling HTE with limited data: a generalisation perspective

Decision modelling aims to estimate the expected net benefits in the target population. Since the trial sample is a selective subset of the target population, how to transport trial estimates remains a key concern. In causal inference, this topic has bee extensively discussed[34–39], using both weighting and regression-based methods. There also exists a class of 'doubly-robust' methods that combines weighting and regression adjustment to improve the robustness and efficiency of the estimator[40–43]. Since the accurate use of weighting methods requires fine-grained target population characteristics, which are rarely available, we focus on outcome regression methods.

Below, we formalize the effect estimation problem and discuss the implications of constant-effect assumption for decision modelling. We then describe the outcome regression methods for generalising results from trials to broader populations under HTE. Finally, we highlight the pitfalls of using highly flexible models when extrapolation is required.

Notation and constant effect model

Here, we formalize the relative effect estimation problem in a trial using the potential outcome framework[44,45]. Suppose that in a trial of size n, each individual's outcome, covariates and treatment assignment indicator can be represented by the triplet $(Y_i, \boldsymbol{X}_i, T_i)$. Let $Y_i(0), Y_i(1)$ denote the two potential outcomes for individual i if they had been treated $(i = 1)$ or untreated $(i = 0)$. The typical target estimand in a trial is the SATE:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)).$$

SATE is the average of individual-level treatment effect at the trial-level. If we assume that the treatment effect is constant across the entire target population, then SATE is unbiased for PATE and can estimate the population-wide benefits in our decision model.

In this constant effect framework, the outcome for individual i can be modelled using generalized linear models (GLM) with or without covariates:

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta^u T_i \tag{3}$$

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta^a T_i + \boldsymbol{\beta} \boldsymbol{X}_i \tag{4}$$

where $g(\cdot)$ is an appropriate link function, β_0 denotes the baseline parameter, β^u and β^a are the coefficient for treatment effects in the marginal and covariate adjusted model, and $\boldsymbol{\beta}$ denotes the prognostic effects of covariates.

For non-collapsible link functions such as the logit, only β^u in (3) targets the SATE. To obtain SATE from (4), marginalization over covariate distributions in the sample IPD is needed[46,47]. The SATE estimate, combined with the estimation for baseline outcome risks from external sources, can then inform the decision model.

Outcome regression for generalization in presence of HTE

If the constant relative effect assumption is invalid, generalizing the trial results requires modellers to consider (i) how covariates modify the treatment effects and (ii) differences in the distributions of effect-modifiers between the trial and target population. Outcome regression methods address both concerns directly: first, by fitting a model targeting the conditional mean of the outcome given treatment and covariates, $\mathrm{E}[Y \mid T, \boldsymbol{X}]$, we can predict outcomes for individuals in the target population under both the treatment and control by plugging in their covariate values. The population-average quantities are then calculated by averaging the individual-level outcomes[48].

A basic approach to modelling HTE is the linear interaction model:

$$g(\mathbb{E}[Y_i]) = \beta_0 + (\beta^{trt} + \boldsymbol{\beta}^{EM} \boldsymbol{X}_i) T_i + \boldsymbol{\beta} \boldsymbol{X}_i \tag{5}$$

where $\boldsymbol{\beta}^{EM}$ captures the effect modification by covariates $\boldsymbol{X}_i$.

While the linear interaction model often tests for HTE, the linear structure can lead to model misspecification as the true relationship is unlikely to be linear. Thus, modern HTE modelling approaches employ more flexible functional forms that adaptively model relationships between covariates and treatment effects to mitigate the chance of model misspecification. Bayesian non-parametric approaches are effective in handling numerous potential effect modifiers. For example, Bayesian additive regression trees (BART)[16,49,50], have demonstrated strong performance in simulation studies[51]. For scenarios with a small number of effect modifiers, splines and generalized additive models can efficiently capture the relationship of interest as they model effect-modifications as smooth, continuous functions of covariates. These models can also be improved by selecting the number and the placements of knots using subject matter expertise. Finally, additional flexibility can be introduced by using separate models for treated and untreated groups, rather than fitting a single outcome regression model[52].

Modelling HTE with limited population inclusion

Reliable estimation of HTEs over the entire covariate space is inherently challenging with limited data. This is particularly relevant for analysis based on efficacy trials, since extrapolation beyond the trial population is required for generalisation. Flexible models, like Bayesian non-parametric approaches, are prone to over-fitting the outcome surface in the data. Naively extrapolating the in-sample outcome-covariate relationship risks making misleading predictions with unwarranted certainty[16]. The spline models we consider in this manuscript also carry this risk: as inherently local methods, their extrapolation is determined by the fit near the boundary, the degree of splines, and smoothness penalties. It is also difficult to assess the correctness of the extrapolation without knowing the *true* outcome surfaces.

It is therefore important to regularize extrapolations to align with existing knowledge of the outcome or treatment effect surfaces. In spline models, constraints can be imposed on smoothness or the overall shape of the curve[53]. Within a Bayesian framework, structured priors provide a principled way to incorporate expert knowledge as shape or smoothness constraints, thereby steering extrapolation behaviours[54]. For novel interventions where little is known about the treatment-effect surface, fitting the outcome surface separately by treatment arms could be a useful alternative.

Once the model is fitted, it can be used to predict outcome probabilities in the target population, informing the treatment effect parameters in downstream decision modelling. Under a fully Bayesian framework, parametric uncertainty in both treatment effects and outcome risks naturally propagates to the predictive distributions of population-average INMB.

We hypothesize that adaptive HTE modelling with appropriate shrinkage will perform well across varying levels of treatment effect heterogeneity. When true heterogeneity is minimal, the model can shrink toward a constant-effect representation; when heterogeneity is present, it can flexibly capture complex outcome surfaces without misrepresenting treatment effects. We explore this in the next section through a simulation study based on the three-state model.

Simulation study

In this exploratory simulation study, we evaluate how different outcome modelling approaches impact predictions of population-level INMB[21] when evidence for HTE is uncertain and limited. This section is structured using the ADEMP framework (Aim, Data-generating mechanisms, Estimand, Methods, and Performance measures)[55].

Aim

Our aim is to assess whether modelling HTE is warranted when the evidence of heterogeneity is uncertain. Specifically, we evaluate whether modelling uncertain HTE improves accuracy of the population-level INMB and improves the representation of decision uncertainty compared to simpler models.

Data-generating mechanisms

The data-generating process consists of two steps: simulating outcome data from a hypothetical clinical trials, and integrating these data into a three-state STM. We firstly describe the trial data generation, followed by the STM setup.

Trial sample formation

The trial is assumed to collet binary outcomes (Y_i) that informs the individual probability of transitions within the three-state STM (p_i). We assume that treatment and age of the patient jointly determine the individual-level transition probabilities. Specifically, using logistic regression,

$$\text{logit}(p_i) = f(age_i) + g(age_i) \times T_i$$

where, $f(\cdot)$ defines the age-specific control outcome ('control surface'), and $g(\cdot)$ the treatment effect ('treatment effect surface'). We consider twelve scenarios by varying the shapes of these two surfaces, summarised in Table 1. The parametrisations are chosen to yield plausible transition probabilities, with the exact expressions for $f(\cdot)$ and $g(\cdot)$ provided in the supplementary material.

The age distribution in the trial is sampled from an age grid. To evaluate how extrapolation impacts predictions, we consider two age sampling approaches: a 'limited' and an 'extended'

Scenario	Control Surface	Treatment Effect Surface
1	Non-linear increasing	Constant
2		Non-linear increasing
3		Non-linear decreasing
4		Non-monotonic
5	Non-linear decreasing	Constant
6		Non-linear increasing
7		Non-linear decreasing
8		Non-monotonic
9	Non-monotonic	Constant
10		Non-linear increasing
11		Non-linear decreasing
12		Non-monotonic

Table 1. Data generating processes separated by different control and treatment effect surfaces on the logit scale

scenario. In the limited scenario, we uniformly sample 500 individuals aged between 40 and 60 per arm. For the extended scenario, we further sample 250 individuals per arm based on a normal distribution centred at 70 with a standard deviation of 5. Full code is in the supplementary materials. For each scenario and sampling strategy, we generate 1,000 Monte Carlo trial-data replicates, forming the IPD from which we estimate transition probabilities.

Decision model setup

We now detail the three-state STM used to translate trial-based estimates into long-term health economic outcomes. The STM comprises three mutually exclusive and exhaustive states: ‘Healthy’, ‘Sick’, and ‘Dead’, Figure 1. All individuals start in the ‘Healthy’ state, while ‘Dead’ is an absorbing state. Transitions between states are governed by an individual-specific matrix $\boldsymbol{P}_i$:

$$\boldsymbol{P}_i = \begin{pmatrix} p_i & 1-p_i & 0 \\ 0 & \frac{p_i}{2} & 1-\frac{p_i}{2} \\ 0 & 0 & 1 \end{pmatrix}, \tag{6}$$

where p_i is the individual-level transition probability estimated from the trial data. To explore the impact of heterogeneity, we artificially related both transition probabilities to p_i.

We simulate the STM over a 40-year time horizon using annual cycles. The proportion of individuals in each state is tracked over time. Each health state s is assigned a fixed cost c_s and a health value e_s. Total costs and effects are calculated by aggregating these over time according

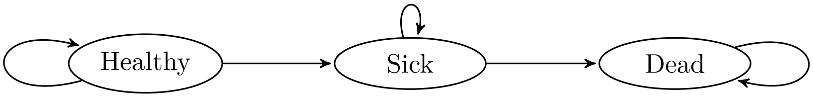

Figure 1. Structure of the three-state Markov state-transition model (STM).

to time spent in each state. All outcomes are discounted at an annual rate of 3%. Finally, we compute the population-level incremental health benefits (Δ_e) and incremental costs (Δ_c). Given a WTP threshold at k, Δ_e and Δ_c can be combined to compute the INMB according to Equation ((2)).

Estimand

The estimand of interest is the INMB in the target population, denoted as INMB^{true}. To derive INMB^{true}, we evaluate the incremental costs and effects from our individual-level STM using 41 age cohorts (ages 40–80) and the exact transition probabilities given by the true outcome surfaces. The resulting age-specific cumulative costs and effects are aggregated using the target population age distribution with 60% of the population uniformly distributed across ages 40–60 and 40% across ages 61–80, following the extended sampling distribution described above. The aggregated population-level costs and health benefits are combined to compute INMB^{true}.

Methods

To compute INMB, we employ a multi-cohort STM with five age cohorts (e.g., 41–50, 51–60, ...) to approximate the individual-level STM. This substantially reduces computational burden. We selected this five-cohort approximation by comparing the predictions generated from models with 5, 9, and 41 cohorts, confirming that the five-cohort approximation is sufficiently accurate.

To estimate the transition probability matrix, we use model-based predictions from included models. From each model, transition probability inputs are random variables from the Bayesian model fitting, allowing for PSA. Specifically, posterior draws of the transition probabilities are obtained by applying the inverse-logit transformation to draws from the posterior distribution of the linear predictor, thereby capturing parameter uncertainty in the outcome mean.

To explore how HTE considerations in outcome models influence decision modelling, we consider two outcome models that ignore HTE, a marginal model (Equation (3)) and a covariate adjusted model (Equation (4)). We then compare the ‘naive’ linear approach to modelling HTE (Equation (5)), and two spline-based models designed specifically for uncertain HTEs.

In spline-based models, we fit the control and treatment-effect surfaces separately to reflect scenarios where prior clinical knowledge may inform the relationship between the covariates and the outcome but offers limited guidance on the treatment-effect relationship. To mitigate spurious extrapolation from limited trial data, we employ two distinct structured priors on spline coefficients for regularisation, resulting in a 'monotonic spline' and an 'unrestricted spline'. Broadly, priors in the monotonic spline model constrain the fitted outcome surface to be monotonic; while the random walk priors in the unrestricted spline model regularise the smoothness. The model is termed 'unrestricted' because it imposes no constraints on monotonicity, shape, or upper bounds in the extrapolated regions. Prior specifications are detailed in the supplementary materials.

Performance comparisons

We evaluate the performance of the models using two complementary metrics. One assesses how the estimated INMB informs decision-making while the other focuses on predictive accuracy.

We first consider the probability of cost-effectiveness, P_{CE}, derived from cost-effectiveness acceptability curves (CEACs). In our primary analysis, we compute P_{CE} at a fixed WTP threshold of £15,000 per QALY. We also explore how P_{CE} evolves as WTP threshold varies from £0 to £30,000 for each modelling approach across simulated data replicates. Because the true population-average INMB (INMB^{true}) is known from the data-generating process, P_{CE} can be interpreted with reference to the 'correct' decision. Ideally, well-calibrated models yield $P_{CE} > 50\%$ when INMB$^{true} > 0$, and $P_{CE} < 50\%$ when INMB$^{true} < 0$. Perfect calibration may not be possible due to limited data and extrapolation uncertainty, but reliable models should still produce decision probabilities that are directionally correct—avoiding high P_{CE} when INMB^{true} barely exceeds zero. Overconfident predictions can mislead decision-makers, so assessing P_{CE} is essential for evaluating practical utility.

While P_{CE} offers insights into decision relevance, purely predictive measures can support statistical model selection. We consider two measures, the log predictive density (LPD) and the posterior predictive mean INMB, both calculated across all 1,000 Monte Carlo replications. The LPD metric quantifies how well the model-based posterior predictions recover INMB^{true}. Conceptually, it is adapted from the predictive measure discussed by Gelman et al.[56], measuring the divergence between the predictive distribution and the truth—a point mass at INMB^{true}. Operationally, we compute the posterior predictive INMB for each model, approximate the density via a kernel density estimator, and evaluate the log probability density at INMB^{true}. As LPD values can be difficult to interpret, we also examine the posterior predictive mean INMB.

Plotting this mean against INMB^{true} across replicates provides a more intuitive assessment of predictive performances.

All simulations were conducted in R version 4.3.1, with scripts made fully available in the supplementary materials.

Results

Before evaluating model performance using the proposed performance measures, we performed an exploratory analysis comparing how outcome surface fits to cost-effectiveness plane predictions for a typical scenario (see supplementary materials). Our main results focus on the 'limited' data scenario, which poses a significant challenge for extrapolation and uncertainty quantification. Results for the 'extended' scenario, which showed similar trends but with attenuated differences between models, are provided in the supplementary materials.

Probability of cost-effectiveness

Figure 2 shows the probability of cost-effectiveness for a WTP of £15000/QALY for the limited scenarios, with scenarios 1,2,5,6,9 and 10 on the left, associated with a positive INMB^{true} and scenarios 3,4,7,8,11 and 12 on the right, associated with a negative INMB^{true}. Thus, we expect the probability of cost-effectiveness to be greater than 50% on the left and less than 50% on the right.

When $\text{INMB}^{true} > 0$ (left), all methods assign relatively high P_{CE} values, typically between 70% and 90%. There are minimal differences in the average P_{CE} across modelling approaches. Both spline models perform well even when the underlying treatment effect is constant (scenarios 1,5 and 9) as the structured Bayesian priors effectively shrink interaction effects towards zero when the when data do not support substantial HTE. The unrestricted spline model tend to be the least confident under limited covariate coverage, but still assigns P_{CE} between 70% and 80%. Importantly, the P_{CE} values have limited correlation with the magnitude of the true INMB, highlighting the challenge of producing well-calibrated predictions based on limited trial data.

Larger differences appear when the intervention is not cost-effective ($\text{INMB}^{true} < 0$). Here, approaches that ignore HTE or model it with a simple linear interaction term appear vastly over-confident, especially in scenario 3, 7 and 11, where the true treatment effect declines with age. These models predict an average P_{CE} of around 80%, failing to capture the reduced treatment benefit in the extrapolated age range. In contrast, both spline models—while still predicting probabilities slightly above 50% due to limited data—express substantially greater decision uncertainty. This dampened confidence level more appropriately signals the ambiguity

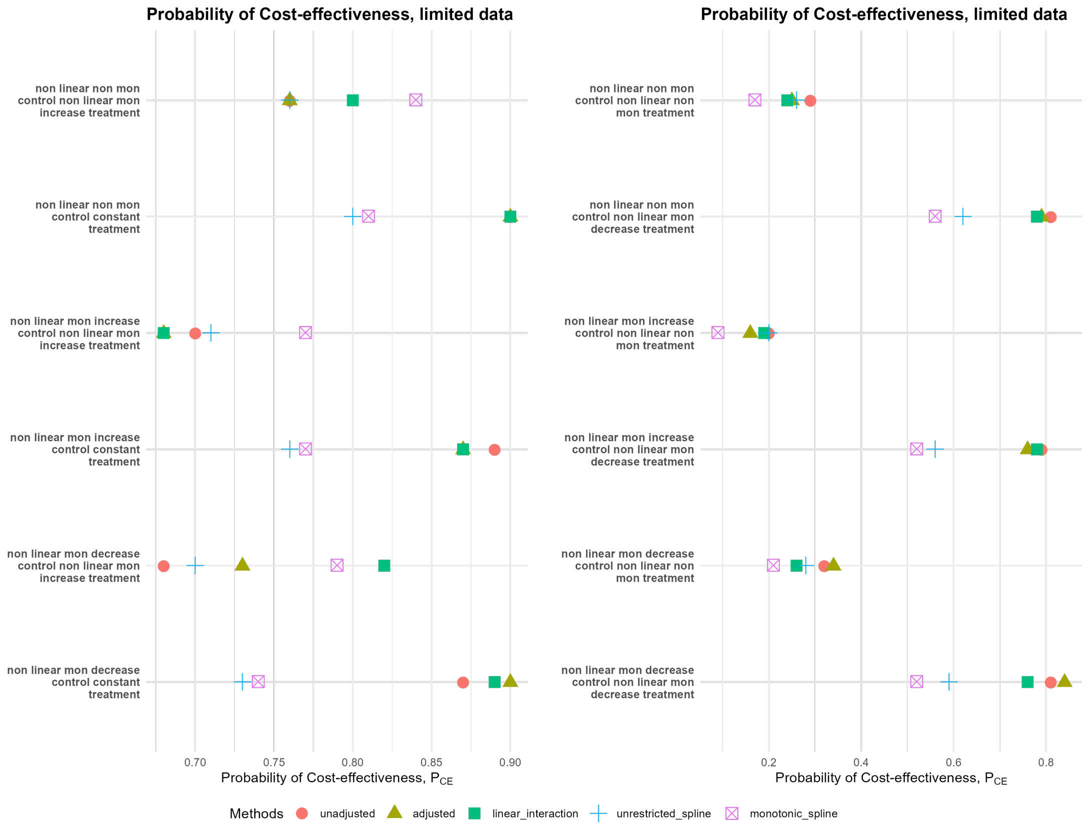

Figure 2. Probability of cost-effectiveness for all outcome modelling approaches across all scenarios under limited covariate coverage.

introduced by uncertain HTE and sparse covariate coverage. From a decision-making standpoint, while the results do not firmly conclude rejection, this reduced confidence might give decision-makers pause.

Figure 3 displays the average CEAC across WTP thresholds from 0 to £30000, across all models and scenarios. One key metric is the predicted probability at the true decision threshold (red dashed line, where INMB$^{true} = 0$), which should be around 50% for well-calibrated predictions. There are key differences in the CEAC at the true decision threshold. In many scenarios, spline models with structured priors yield P_{CE} values considerably closer to 50% than other approaches. These models also tend to assign lower probabilities overall, with more gradually rising CEACs, indicating greater uncertainty. By contrast, other methods frequently predict P_{CE} values substantially above 50% at the decision threshold, indicating overconfidence.

Overall, these findings suggest that modelling HTE with appropriate regularization improves decision-making by better representing the uncertainty in population-level cost-effectiveness. While none of the models produce fully calibrated predictions under limited data, the added caution from regularised splines helps avoid overly confident and potentially misleading policy recommendations.

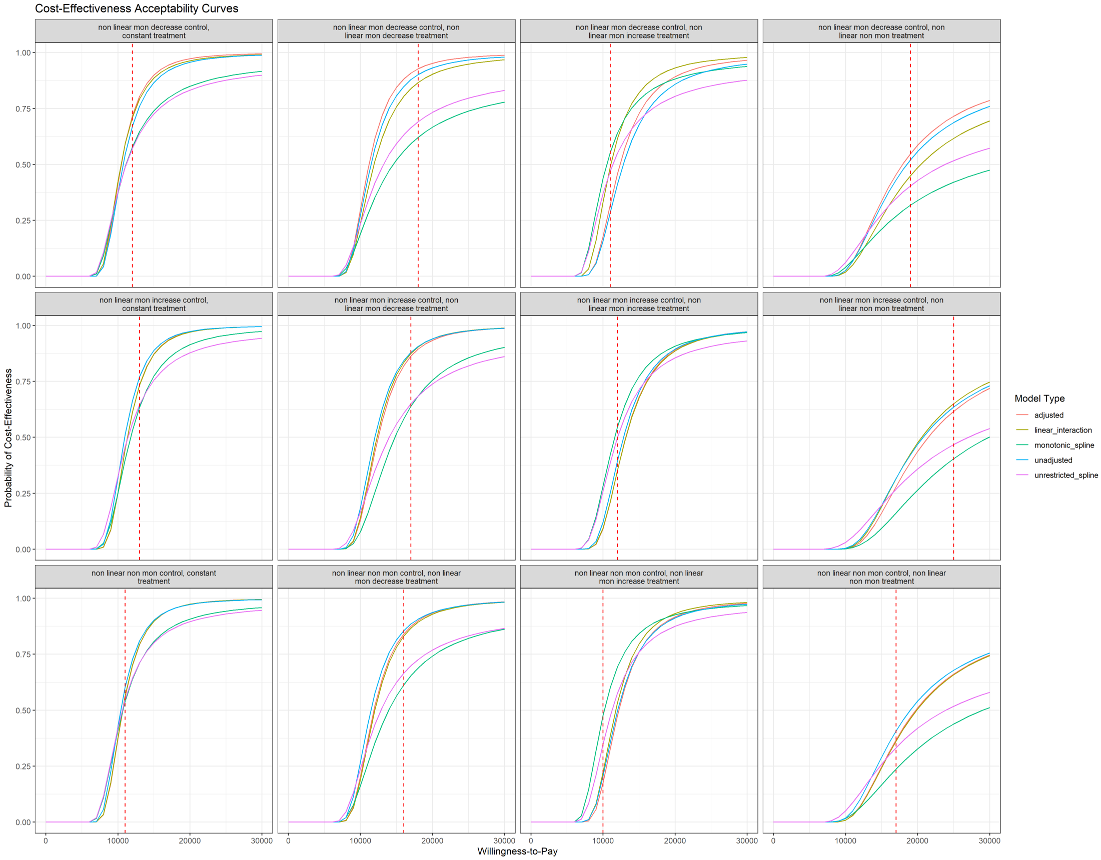

Figure 3. Cost-effectiveness acceptability curve for all outcome modelling approaches based on average probabilities across all scenarios under limited covariate coverage. The red dashed line indicates the willingness to pay value where the true INMB first becomes positive. Overall, the two spline approaches for modelling heterogenous treatment effects exhibit the least confidence across all scenarios

Predictive accuracy metrics

We assessed the distribution of posterior predictive mean INMB and the average LPD of INMB^{true} across 1000 Monte Carlo replications, visualized in Figures 4 and 5, respectively. Numerical results are provided in the supplementary materials.

Figure 4 illustrates the difficulties associated with point prediction under limited covariate coverage. While the overall mean of the posterior predictive INMB across all Monte Carlo replications can closely align with the true value (red triangle), there is a lot of variation in the mean INMB based on the specific dataset, evidenced by the wide 95% uncertainty intervals. This signifies that an INMB estimate derived from any given trial sample could diverge considerably from the truth. Therefore, achieving reliable point prediction for population-average INMB appears highly challenging, if not impossible, based on samples with limited covariate coverage. This has important implications for decision making as the optimal decision is identified by the population-average INMB.

Turning to the LPD results in Figure 5, neither the average LPD scores nor the overlapping uncertainty intervals identify a consistently superior modelling strategy for recovering INMB^{true}. However, the width of these intervals, reflecting the empirical variability of LPD scores across replications, highlights differences in predictive stability. The regularized spline models exhibit narrower intervals compared to the other approaches in most of the scenarios, indicating their predictive performance is less sensitive to the specific trial sample. This greater stability might stem from the spline models making more diffuse posterior predictions for the INMB.

Ultimately, this predictive accuracy analysis underscores the inherent difficulty, and perhaps impracticality, of precisely estimating population-average cost-effectiveness from a single trial with limited covariate coverage. By focusing on methods that appropriately reflect the uncertainty in the extrapolation from limited data, we can provide a more realistic basis for decision-making.

Discussion

This manuscript investigates how outcome modelling approaches that account for uncertain HTE impact the estimation of population-level cost-effectiveness within trial-based decision modelling contexts. We focus on a modelling strategy that integrates IPD from an efficacy trial with a STM. Our simulation explored a simplified scenario with a binary outcome and a single continuous covariate. We varied the underlying shapes of outcome and treatment effect surfaces and evaluated model performance across twelve distinct scenarios.

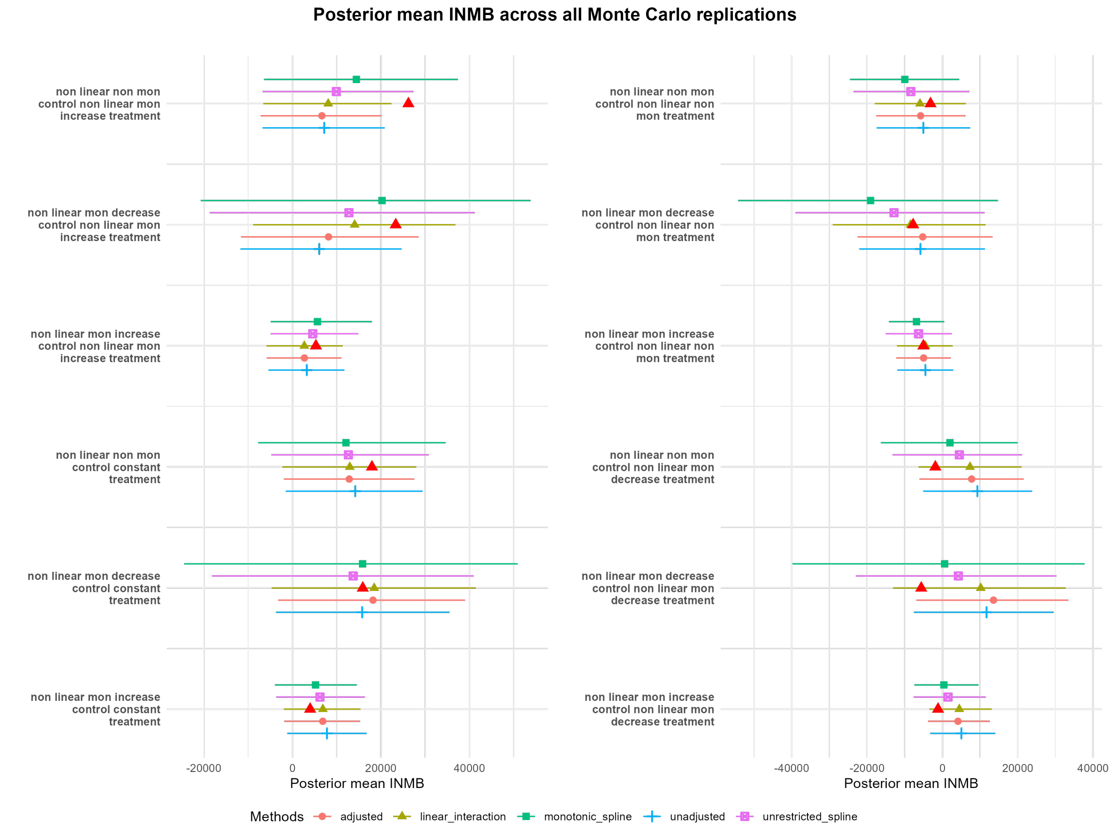

Figure 4. Posterior predictive mean INMB across all Monte Carlo replications with the true INMB overlaid as a red triangle. The point range plots for each method indicate the overall mean and the empirical 95% intervals of the posterior predictive mean. None of the methods provide satisfactory predictive performances, as predictions from any arbitrary sample can be rather different from the true value. The overall mean can be close to the truth in some scenarios. But that is irrelevant to real-world analysis where the prediction is always conditional on the sample at hand.

Our simulation study demonstrates that flexible models incorporating HTE, with structured prior regularization, offer comparable predictive performance across all scenarios while offering superior representations of decision uncertainty. Crucially, they mitigate the substantial overconfidence in models ignoring HTE or extrapolation uncertainty, particularly when true treatment effects decrease across the covariate range. This preliminary finding challenges the conventional wisdom that ignoring HTE, under uncertain supporting evidence, can be justifiable from a conservative stance towards decision making.

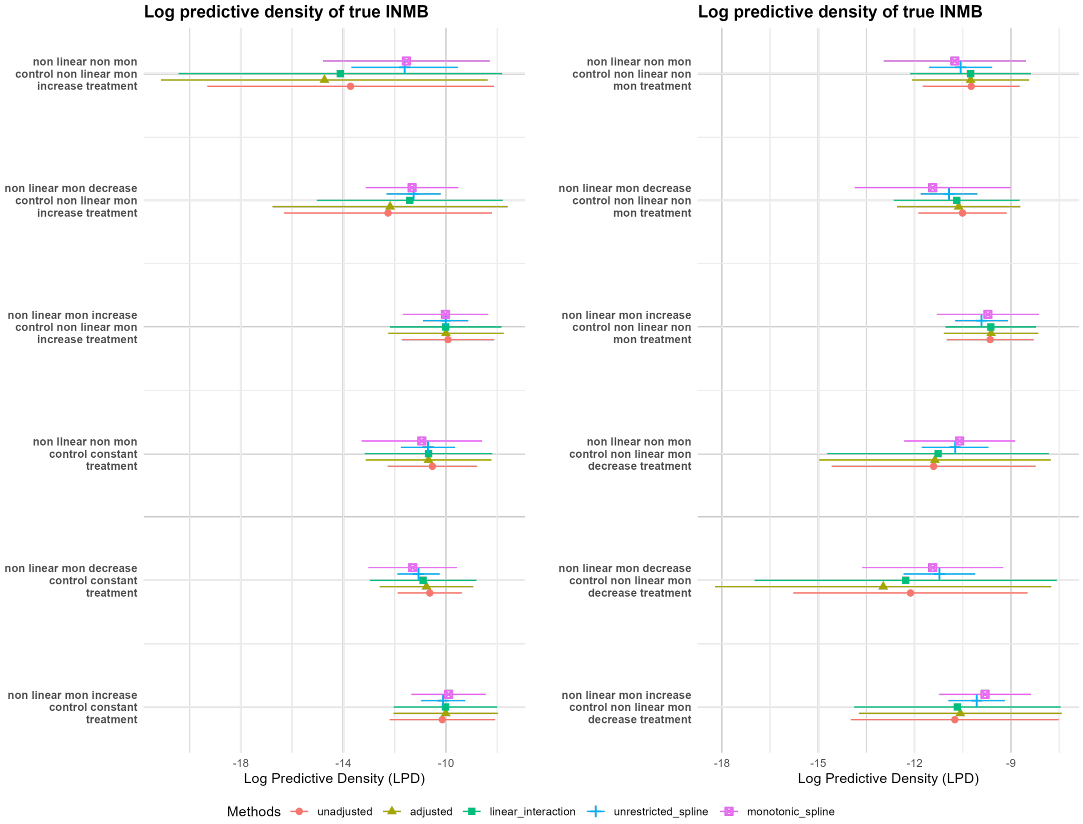

Figure 5. Log predictive density of the true INMB across scenarios under limited covariate coverage. The points show the grand mean across all Monte Carlo replications while the intervals indicate 2 times the empirical standard errors from the mean. Despite the differences in the overall mean, the predictions can vary substantially based on the sample of analysis. In general, none of the methods stand out in this model comparison metric.

As this simulation is illustrative, we do not advocate for a specific outcome model. Instead, model complexity should be guided by substantive knowledge of the disease process and not constrained by sample size. While complex models fitted to limited data naturally yield greater parametric uncertainty, this is not inherently problematic. Quantifying this uncertainty transparently is vital for decision-making and enables formal Value of Information analysis to guide research prioritization and trial planning[57–62].

A key limitation the simplified data generating process in our simulation. Real-world health economic evaluations often involve trials with numerous prognostic factors and potential effect modifiers, increasing the risk of outcome model misspecification, especially when extrapolating beyond the observed covariate range. Previous research has combined Bayesian non-parametric models with penalized splines to extrapolate into regions with limited covariate overlap[63], although not within health economics. While our study does directly address more complex scenarios, the structured priors studied here provide a practical means of guiding extrapolations in data-sparse regions. These regularization strategies may thus remain valuable in more realistic modelling scenarios in future research.

Despite these limitations, our study offers practical insights, particularly for the increasingly common scenarios where HTE is plausible but trial data provides limited covariate coverage and statistical power to characterize it. Our findings suggest that adopting flexible, regularized models that account for HTE is advantageous compared with defaulting to simpler, constant-effect assumptions. These flexible models help avoid potentially misleading, overly confident conclusions about cost-effectiveness. Furthermore, our study highlights the challenges of predicting population-level INMB based on a single RCT, even when the trial-based analysis is further integrated with a decision analytic model. Future research could explore hybrid approaches that combine RCTs with observational data to estimate HTEs, which may improve the accuracy of population-level cost-effectiveness predictions when integrated into decision models.

References

1. Goodacre S and McCabe C. An introduction to economic evaluation. *Emergency medicine journal: EMJ* 2002; 19(3): 198.
2. Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoeconomics* 2008; 26: 799–806.
3. Kuntz KM and Goldie SJ. Assessing the sensitivity of decision-analytic results to unobserved markers of risk: defining the effects of heterogeneity bias. *Medical decision making* 2002; 22(3): 218–227.
4. Zaric GS. The impact of ignoring population heterogeneity when markov models are used in cost-effectiveness analysis. *Medical Decision Making* 2003; 23(5): 379–386.
5. Elbasha EH and Chhatwal J. Characterizing heterogeneity bias in cohort-based models. *PharmacoEconomics* 2015; 33: 857–865.
6. van Rosmalen J, Zauber AG, van Ballegooijen M et al. Multicohort models in cost-effectiveness analysis. *Medical Decision Making* 2013; 33(3): 407–414.
7. 5 The reference case — Guide to the methods of technology appraisal 2013 — Guidance — NICE. https://www.nice.org.uk/process/pmg9/chapter/the-reference-case, 2013.

8. Sadique Z, Grieve R, Diaz-Ordaz K et al. A machine-learning approach for estimating subgroup-and individual-level treatment effects: an illustration using the 65 trial. *Medical Decision Making* 2022; 42(7): 923–936.
9. Padula WV, Kreif N, Vanness DJ et al. Machine learning methods in health economics and outcomes research—the palisade checklist: a good practices report of an ispor task force. *Value in health* 2022; 25(7): 1063–1080.
10. Bonander C and Svensson M. Using causal forests to assess heterogeneity in cost-effectiveness analysis. *Health Economics* 2021; 30(8): 1818–1832.
11. Dias S, Welton NJ, Sutton AJ et al. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Medical Decision Making* 2013; 33(5): 641–656.
12. Tierney JF, Vale C, Riley R et al. Individual participant data (ipd) meta-analyses of randomised controlled trials: guidance on their use. *PLoS medicine* 2015; 12(7): e1001855.
13. Riley RD, Dias S, Donegan S et al. Using individual participant data to improve network meta-analysis projects. *BMJ evidence-based medicine* 2023; 28(3): 197–203.
14. Schwartz D and Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases* 1967; 20(8): 637–648. DOI:10.1016/0021-9681(67)90041-0.
15. Shields GE, Wilberforce M, Clarkson P et al. Factors limiting subgroup analysis in cost-effectiveness analysis and a call for transparency. *Pharmacoeconomics* 2022; 40(2): 149–156.
16. Hill JL. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 2011; 20(1): 217–240.
17. Steuten LMG, van de Wetering G, Groothuis-Oudshoorn K et al. A systematic and critical review of the evolving methods and applications of value of information in academia and practice. *Pharmacoeconomics* 2013; 31(1): 25–48. DOI:10.1007/s40273-012-0008-3.
18. Heath A, Kunst N and Jackson C. *Value of Information for Healthcare Decision-Making.* CRC Press, 2024.
19. Drummond MF, Sculpher MJ, Claxton K et al. *Methods for the economic evaluation of health care programmes.* Oxford university press, 2015.
20. National Institute for Health and Care Excellence. *NICE health technology evaluations: the manual*, 2023. URL https://www.nice.org.uk/process/pmg36. Published 31 January 2022. Last updated October 2023. Process guide PMG36. Available from: https://www.nice.org.uk/process/pmg36.
21. Stinnett AA and Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical decision making* 1998; 18(2_suppl): S68–S80.
22. Briggs A, Sculpher M and Claxton K. *Decision modelling for health economic evaluation.* Oup Oxford, 2006.

23. O'Hagan A, McCabe C, Akehurst R et al. Incorporation of uncertainty in health economic modelling studies. *Pharmacoeconomics* 2005; 23: 529–536.
24. Claxton K, Sculpher M, McCabe C et al. Probabilistic sensitivity analysis for nice technology assessment: not an optional extra. *Health economics* 2005; 14(4): 339–347.
25. Baio G and Dawid AP. Probabilistic sensitivity analysis in health economics. *Statistical methods in medical research* 2015; 24(6): 615–634.
26. Baio G. *Bayesian methods in health economics.* CRC Press, 2012.
27. Karnon J, Stahl J, Brennan A et al. Modeling using discrete event simulation: a report of the ispor-smdm modeling good research practices task force–4. *Medical decision making* 2012; 32(5): 701–711.
28. Davis S, Stevenson M, Tappenden P et al. Nice dsu technical support document 15: Cost-effectiveness modelling using patient-level simulation. *Rep BY Decis Support UNIT* 2014; .
29. Espinoza MA, Manca A, Claxton K et al. The value of heterogeneity for cost-effectiveness subgroup analysis: Conceptual framework and application. *Med Decis Making* 2014; 34(8): 951–964. DOI: 10.1177/0272989X14538705.
30. Basu A and Meltzer D. Value of information on preference heterogeneity and individualized care. *Medical Decision Making* 2007; 27(2): 112–127.
31. Dias S, Sutton A, Welton N et al. *NICE DSU Technical Support Document 3: Heterogeneity: Subgroups, Meta-Regression, Bias and Bias-Adjustment.* National Institute for Health and Clinical Excellence, 2011.
32. Turner AJ, Sammon C, Latimer N et al. Transporting comparative effectiveness evidence between countries: considerations for health technology assessments. *Pharmacoeconomics* 2024; 42(2): 165–176.
33. CADTH. Guidance for reporting real-world evidence. Technical report, Canada's Drug and Health Technology Agency (CDA-AMC), 2023. URL https://www.cda-amc.ca/sites/default/files/RWE/MG0020/MG0020-RWE-Guidance-Report-Secured.pdf. Accessed: February 26, 2025.
34. Kennedy-Martin T, Curtis S, Faries D et al. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015; 16: 1–14.
35. Imai K, King G and Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)* 2008; 171(2): 481–502.
36. Cole SR and Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology* 2010; 172(1): 107–115.

37. Cook TD, Campbell DT and Shadish W. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Boston, MA, 2002.
38. Stuart EA, Ackerman B and Westreich D. Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on social work practice* 2018; 28(5): 532–537.
39. Degtiar I and Rose S. A review of generalizability and transportability. *Annual Review of Statistics and Its Application* 2023; 10: 501–524.
40. JM R, A R and LP Z. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; 89(427): 846–866. DOI:10.1080/01621459.1994.10476818.
41. JDY K and JL S. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007; 22(4): 523–539. DOI:10.1214/07-STS227.
42. H B and JM R. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; 61(4): 962–972. DOI:10.1111/j.1541-0420.2005.00377.x.
43. J H. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998; 66(2): 315–331. DOI:10.2307/2998560.
44. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 1974; 66(5): 688.
45. Splawa-Neyman J, Dabrowska DM and Speed TP. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science* 1990; : 465–472.
46. Greenland S, Pearl J and Robins JM. Confounding and collapsibility in causal inference. *Statistical science* 1999; 14(1): 29–46.
47. Daniel R, Zhang J and Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal* 2021; 63(3): 528–557.
48. Remiro-Azócar A, Heath A and Baio G. Parametric g-computation for compatible indirect treatment comparisons with limited individual patient data. *Research synthesis methods* 2022; 13(6): 716–744.
49. Hahn PR, Murray JS and Carvalho CM. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis* 2020; 15(3): 965–1056.
50. Hill J, Linero A and Murray J. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application* 2020; 7: 251–278.
51. Dorie V, Hill J, Shalit U et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science* 2019; 34(1): 43 – 68.

52. Künzel SR, Sekhon JS, Bickel PJ et al. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 2019; 116(10): 4156–4165.

53. Pya N and Wood SN. Shape constrained additive models. *Statistics and computing* 2015; 25: 543–559.

54. Gao Y, Kennedy L, Simpson D et al. Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis* 2020; 16(3): 719.

55. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in medicine* 2019; 38(11): 2074–2102.

56. Gelman A, Hwang J and Vehtari A. Understanding predictive information criteria for bayesian models. *Statistics and computing* 2014; 24(6): 997–1016.

57. Ades A, Lu G and Claxton K. Expected value of sample information calculations in medical decision modeling. *Medical decision making* 2004; 24(2): 207–227.

58. Fenwick E, Palmer S, Claxton K et al. An iterative bayesian approach to health technology assessment: application to a policy of preoperative optimization for patients undergoing major elective surgery. *Medical Decision Making* 2006; 26(5): 480–496.

59. Minelli C and Baio G. Value of information: a tool to improve research prioritization and reduce waste. *PLoS medicine* 2015; 12(9): e1001882.

60. Felli JC and Hazen GB. Sensitivity analysis and the expected value of perfect information. *Medical Decision Making* 1998; 18(1): 95–109.

61. Willan AR and Pinto EM. The value of information and optimal clinical trial design. *Statistics in medicine* 2005; 24(12): 1791–1806.

62. McKenna C and Claxton K. Addressing adoption and research design decisions simultaneously: the role of value of sample information analysis. *Medical Decision Making* 2011; 31(6).

63. Nethery RC, Mealli F and Dominici F. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The annals of applied statistics* 2019; 13(2): 1242.

Appendix

Prior specifications for the spline model

- Unrestricted spline model:

$$
\begin{aligned}
\text{Intercept} &\sim \mathcal{N}(0,3) \\
\alpha &\sim \text{Exponential}(1) \\
\gamma_1 &\sim \text{Normal}(0,1) \\
\gamma_j &\sim \text{Normal}(\gamma_{j-1},1) \text{ for } j=2,\ldots,m \\
\sum_{j=1}^{m}\gamma_j &\sim \mathcal{N}(0,0.01\times m) \\
\boldsymbol{s_0} &= \alpha\times\boldsymbol{\gamma} \\
\text{logit}(p_i) &= \text{Intercept} + S_i\cdot\boldsymbol{s_0} \\
Y_i &\sim \text{Bernoulli}(p_i)
\end{aligned} \tag{7}
$$

  where:

  - S_i is the B-spline basis matrix of age_i.
  - The priors on γ_i together with the soft sum to zero constraint effectively implements a random walk prior on the spline coefficients

- Monotonic spline model:

$$
\begin{aligned}
\text{Intercept} &\sim \mathcal{N}(0,3) \\
\alpha &\sim \text{Exponential}(1) \\
\boldsymbol{\gamma} &\sim \text{Logistic}(0,0.5) \\
\boldsymbol{b} &= (0,\gamma_1,\ldots,\gamma_{m-1}) \\
\boldsymbol{c} &= \text{softmax}(\boldsymbol{b}) \\
\boldsymbol{s_0} &= \alpha\times\boldsymbol{c} \\
\text{logit}(p_i) &= \text{Intercept} + S_i\cdot\boldsymbol{s_0} \\
Y_i &\sim \text{Bernoulli}(p_i)
\end{aligned} \tag{8}
$$

  where:

- S_i is the integrated B-spline basis matrix of age_i.
- The softmax refers to the operation:

$$\text{softmax}(b_i) = \frac{e^{b_i}}{\sum_{j=1}^{m} e^{b_j}}$$

Outcome surface fits and cost-effectiveness plane predictions

We perform exploratory analysis investigating how the fit to the outcome surface can impact the fit in the cost-effectiveness plane. Below, for demonstration purposes, we randomly select one result out of 1000 Monte Carlo replications for scenario 7 and plot the fits for both 'limited' and 'extended' scenarios for all outcome modelling approaches. The outcome surface plots for the unadjusted model are omitted as they would simply be a horizontal line. The fit to the cost-effectiveness plane and the fit to the outcome surfaces are presented in a side-by-side manner to aid comparisons.

As shown in Figure 6, even with structured priors, the fit to the outcome surface remains unsatisfactory in the limited scenario. The corresponding predictions in the cost-effectiveness plane are equally unsatisfactory. Only the two spline models produce predictions that cover the ground truth but they achieve this by inflating the uncertainty in the extrapolation of the outcome surfaces. As the extended scenario includes higher data coverage in the entire age range, they clearly exhibit a better fit to the true outcome surface. Consequently, the differences among five modelling approaches in the cost-effectiveness plane appear to be smaller. Importantly, the 'covering' INMB^{true} in the cost-effectiveness plane does not directly translate to good probabilistic predictions of INMB^{true}—the LPD assigned to the true value can still be low.

This basic exploratory analysis demonstrates that it is indeed difficult to provide good approximations to the true outcome surface with sample data that only span part of the target covariate range. Nevertheless, in the presence of HTE, an uncertain model that accounts for HTE might provide better predictions than ones ignoring HTE altogether - in the 'limited scenario', the two spline models exhibit coverage of the true value by 'acknowledging' the large uncertainty in extrapolations.

Simulation results under extended covariate coverage

Additional simulation results

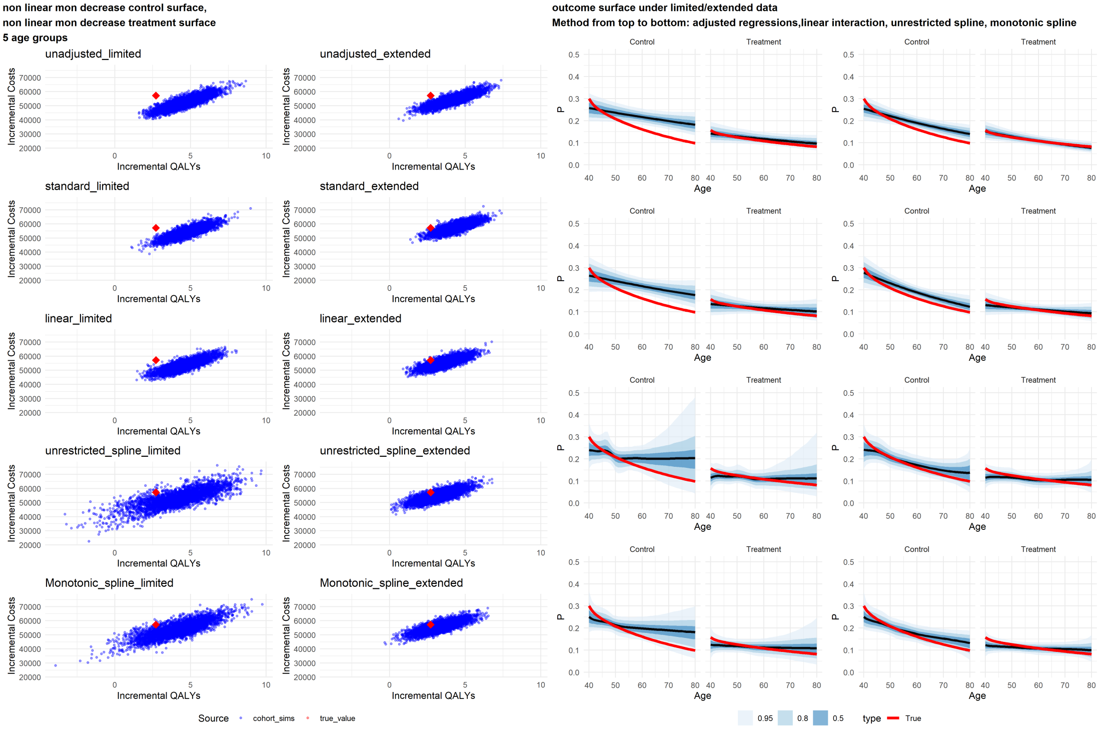

Figure 6. Outcome surface vs cost-effectiveness plane under non-linear decreasing control surface and non-linear decreasing treatment surface. The outcome surface plot for the unadjusted model is omitted here. Except for that, each row of the outcome surface fit on the right correspond to a cost-effectiveness plane on the left. Overall, the fit of the outcome surface does not directly translate to a 'good' coverage on the cost-effectiveness plane.

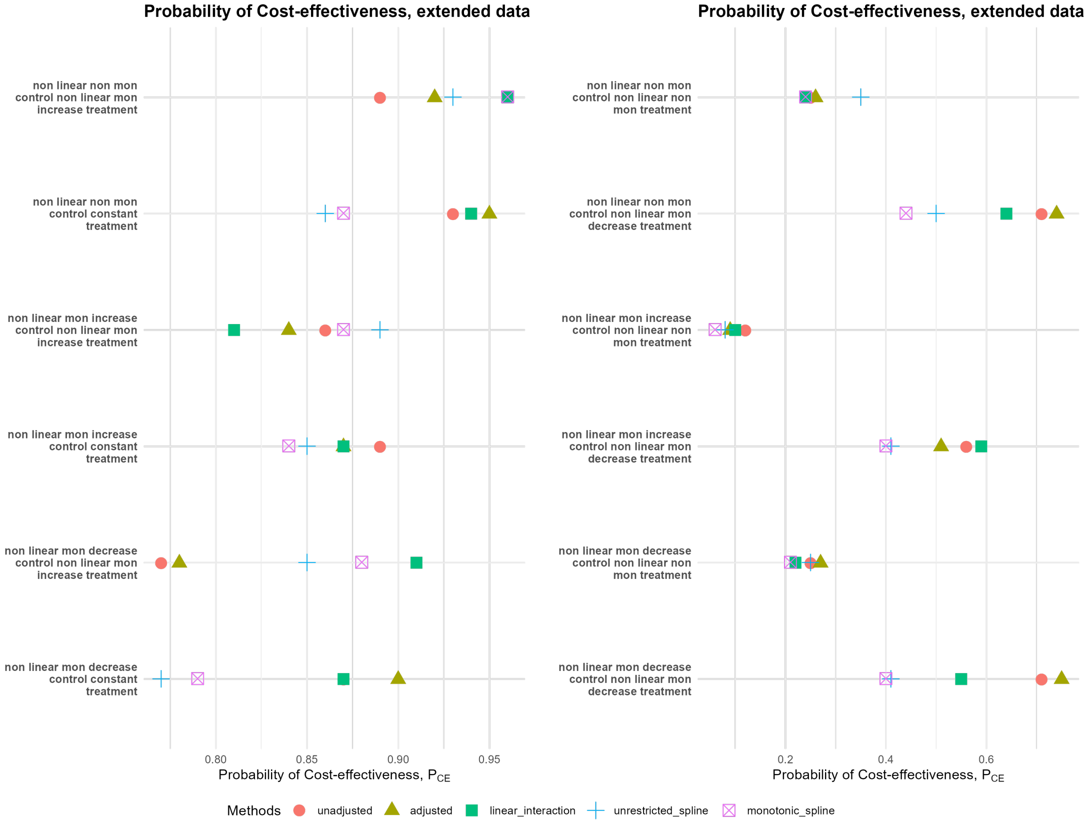

Figure 7. Probability of cost-effectiveness for all outcome modelling approaches across all scenarios under extended covariate coverage.

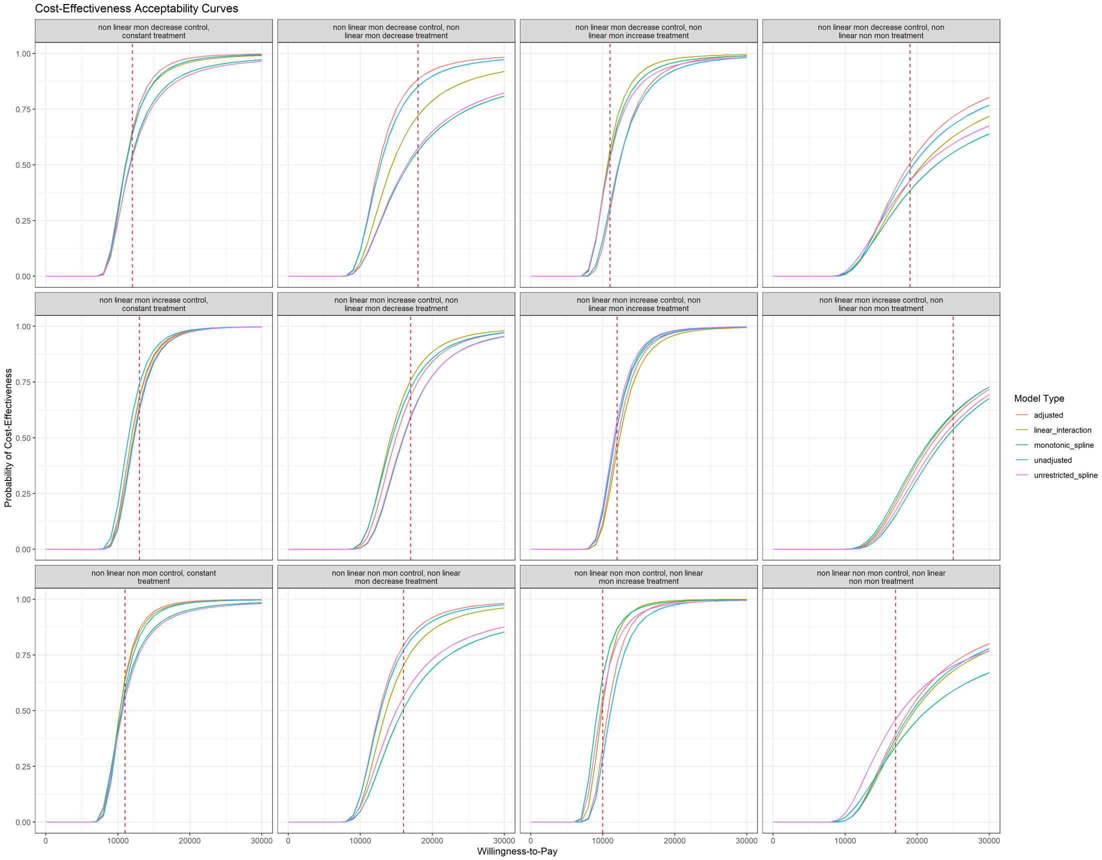

Figure 8. Cost-effectiveness acceptability curve for all outcome modelling approaches based on average probabilities across all scenarios under extended covariate coverage. The red dashed line indicates the willingness to pay value where the true INMB first becomes positive. In general, with extended covariate coverage, there appears to be a convergence trend. But the spline models are still comparatively under-confident

Scenario	Unadjusted	Adjusted	Linear Interaction	Unrestricted Spline	Monotonic Spline
non linear mon increase control constant treatment	-10.14 (1.05)	-10.01 (1.04)	-10.02 (1.03)	-10.11 (0.44)	-9.9 (0.74)
non linear mon decrease control constant treatment	-10.62 (0.64)	-10.76 (0.93)	-10.89 (1.06)	-11.07 (0.42)	-11.3 (0.88)
non linear non mon control constant treatment	-10.53 (0.89)	-10.68 (1.25)	-10.68 (1.27)	-10.7 (0.54)	-10.94 (1.2)
non linear mon increase control non linear mon increase treatment	-9.92 (0.92)	-10 (1.15)	-10.01 (1.11)	-10.01 (0.45)	-10.02 (0.85)
non linear mon decrease control non linear mon increase treatment	-12.26 (2.07)	-12.18 (2.34)	-11.41 (1.85)	-11.26 (0.54)	-11.32 (0.92)
non linear non mon control non linear mon increase treatment	-13.72 (2.85)	-14.74 (3.25)	-14.12 (3.22)	-11.61 (1.06)	-11.54 (1.66)
non linear mon increase control non linear mon decrease treatment	-10.75 (1.65)	-10.58 (1.61)	-10.67 (1.64)	-10.07 (0.45)	-9.81 (0.73)
non linear mon decrease control non linear mon decrease treatment	-12.13 (1.86)	-12.98 (2.66)	-12.28 (2.4)	-11.23 (0.57)	-11.44 (1.12)
non linear non mon control non linear mon decrease treatment	-11.41 (1.62)	-11.37 (1.84)	-11.27 (1.76)	-10.73 (0.53)	-10.6 (0.88)
non linear mon increase control non linear non mon treatment	-9.65 (0.69)	-9.62 (0.75)	-9.63 (0.72)	-9.92 (0.42)	-9.72 (0.81)
non linear mon decrease control non linear non mon treatment	-10.51 (0.7)	-10.63 (0.98)	-10.69 (1)	-10.93 (0.45)	-11.44 (1.24)
non linear non mon control non linear non mon treatment	-10.24 (0.77)	-10.26 (0.93)	-10.26 (0.96)	-10.57 (0.5)	-10.75 (1.13)

Table 2. Average log predictive density of true INMB across all scenarios. The empirical standard errors are shown within the bracket. All models are roughly equivalent based on this scoring rule alone. Meanwhile, approaches that ignore HTE tends to have higher empirical standard errors compared with the two spline models.

Scenario	True INMB	Unadjusted	Adjusted	Linear interaction	Unrestricted spline	Monotonic spline
non linear mon increase control constant treatment	4011	0.89	0.87	0.87	0.76	0.77
non linear mon decrease control constant treatment	15907	0.87	0.9	0.89	0.73	0.74
non linear non mon control constant treatment	17973	0.9	0.9	0.9	0.8	0.81
non linear mon increase control non linear mon increase treatment	5308	0.7	0.68	0.68	0.71	0.77
non linear mon decrease control non linear mon increase treatment	23350	0.68	0.73	0.82	0.7	0.79
non linear non mon control non linear mon increase treatment	26219	0.76	0.76	0.8	0.76	0.84
non linear mon increase control non linear mon decrease treatment	-1150	0.79	0.76	0.78	0.56	0.52
non linear mon decrease control non linear mon decrease treatment	-5610	0.81	0.84	0.76	0.59	0.52
non linear non mon control non linear mon decrease treatment	-1877	0.81	0.79	0.78	0.62	0.56
non linear mon increase control non linear non mon treatment	-5048	0.2	0.16	0.19	0.2	0.09
non linear mon decrease control non linear non mon treatment	-7765	0.32	0.34	0.26	0.28	0.21
non linear non mon control non linear non mon treatment	-3149	0.29	0.25	0.24	0.26	0.17

Table 3. Average acceptance probability across all scenarios. The True INMB is shown in the second column. It appears that predicted acceptance probability tends to be over-confident, especially in scenario 7 and 8. Using priors to steer extrapolation tends to give more 'calibrated; acceptance probability

Scenario	True INMB	Unadjusted	Adjusted	Linear interaction	Unrestricted spline	Monotonic spline
non linear mon increase control constant treatment	4011	0.89	0.87	0.87	0.85	0.84
non linear mon decrease control constant treatment	15907	0.87	0.9	0.87	0.77	0.79
non linear non mon control constant treatment	17974	0.93	0.95	0.94	0.86	0.87
non linear mon increase control non linear mon increase treatment	5308	0.86	0.84	0.81	0.89	0.87
non linear mon decrease control non linear mon increase treatment	23351	0.77	0.78	0.91	0.85	0.88
non linear non mon control non linear mon increase treatment	26219	0.89	0.92	0.96	0.93	0.96
non linear mon increase control non linear mon decrease treatment	-1150	0.56	0.51	0.59	0.41	0.4
non linear mon decrease control non linear mon decrease treatment	-5610	0.71	0.75	0.55	0.41	0.4
non linear non mon control non linear mon decrease treatment	-1878	0.71	0.74	0.64	0.5	0.44
non linear mon increase control non linear non mon treatment	-5049	0.12	0.09	0.1	0.08	0.06
non linear mon decrease control non linear non mon treatment	-7765	0.25	0.27	0.22	0.25	0.21
non linear non mon control non linear non mon treatment	-3149	0.25	0.26	0.24	0.35	0.24

Table 4. Average acceptance probability across scenarios with extended data. Compared to the data with limited range, extending the covariate range with fewer observations does not solve the problem of over-confidence if HTE is ignored