Article:
Jennings, A.C., Rutherford, M.J., Latimer, N.R. orcid.org/0000-0001-5304-5585 et al. (2 more authors) (2024) Perils of randomised controlled trial survival extrapolation assuming treatment effect waning: why the distinction between marginal and conditional estimates matters. Value in Health, 27 (3). pp. 347-355. ISSN 1098-3015

https://doi.org/10.1016/j.jval.2023.12.008

ScienceDirect

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

Methodology

Perils of Randomized Controlled Trial Survival Extrapolation Assuming Treatment Effect Waning: Why the Distinction Between Marginal and Conditional Estimates Matters

Angus C. Jennings, MSc, Mark J. Rutherford, PhD, Nicholas R. Latimer, PhD, Michael J. Sweeting, PhD, Paul C. Lambert, PhD

ABSTRACT

Objectives: A long-term, constant, protective treatment effect is a strong assumption when extrapolating survival beyond clinical trial follow-up; hence, sensitivity to treatment effect waning is commonly assessed for economic evaluations. Forcing a hazard ratio (HR) to 1 does not necessarily estimate loss of individual-level treatment effect accurately because of HR selection bias. A simulation study was designed to explore the behavior of marginal HRs under a waning conditional (individual-level) treatment effect and demonstrate bias in forcing a marginal HR to 1 when the estimand is "survival difference with individual-level waning".

Methods: Data were simulated under 4 parameter combinations (varying prognostic strength of heterogeneity and treatment effect). Time-varying marginal HRs were estimated in scenarios where the true conditional HR attenuated to 1. Restricted mean survival time differences, estimated having constrained the marginal HR to 1, were compared with true values to assess bias induced by marginal constraints.

Results: Under loss of conditional treatment effect, the marginal HR took a value >1 because of covariate imbalances. Constraining this value to 1 lead to restricted mean survival time difference bias of up to 0.8 years (57% increase). Inflation of effect size estimates also increased with the magnitude of initial protective treatment effect.

Conclusions: Important differences exist between survival extrapolations assuming marginal versus conditional treatment effect waning. When a marginal HR is constrained to 1 to assess efficacy under individual-level treatment effect waning, the survival benefits associated with the new treatment will be overestimated, and incremental cost-effectiveness ratios will be underestimated.

Keywords: efficacy waning, hazard ratio noncollapsability, health technology assessment, survival extrapolation, treatment effect waning.

VALUE HEALTH. 2024; 27(3):347–355

Introduction

Extrapolation of a time-to-event treatment effect beyond randomized controlled trial (RCT) follow-up is required to estimate lifetime treatment benefits and is important for accurate health technology assessment (HTA). Hazard ratio (HR) estimates under randomization often assume proportional hazards (PHs)[1]; extrapolating survival based on this implies a constant reduction in instantaneous event rate at all times to a lifetime horizon.

Treatment effect waning assumptions are implemented to estimate long-term treatment gains (eg, in quality-adjusted life-years) assuming treatment effectiveness wanes over a given period or to model the impacts of a treatment stopping rule. Whether of direct interest or used to assess sensitivity to more pessimistic outcomes with a paucity of long-term data, its consideration is suggested by National Institute for Health and Care Excellence (NICE)[2] for HTAs in England. Because treatment discontinuation due to stopping rules and loss of effectiveness happen at an individual level, analyses exploring these issues imply the estimand, "time-to-event or survival difference under individual-level efficacy waning."

Waning assumptions are most commonly reported in oncology[3] and multiple sclerosis[4] HTAs. Further examples exist across vaccination,[5] chronic illness,[6,7] gene therapy,[8] cardiovascular disease,[9] kidney disease,[10–12] and skin conditions.[13,14] Their use is often justified by an insufficiency of evidence to conclude a maintained treatment effect into long-term follow-up or because of the implementation of treatment-stopping rules. For many clinical contexts, efficacy waning may not be relevant, eg, for a curative therapy, or where long-term follow-up data may be sufficient to conclude maintained efficacy. Discussion tends to focus on the time point waning might be applied from and whether it

should be instantaneous or gradual, with discourse around covariate adjustment lacking.

In a time-to-event scenario, treatment effect waning is commonly implemented by assuming a HR approaches 1 from a specified time, often instantly or in predefined steps,[15,16] reflecting NICE recommendations to explore scenarios assuming technology does not provide further benefit after use, as well as more optimistic outcomes eg, gradual diminution over time[2].

The inherent selection bias in HRs with unmodeled frailty, caused by consideration of only the at-risk population for calculations (discussed by Hernán[17]), complicates extrapolations. Systematic post-randomization prognostic factor imbalances will exist between survivors in different treatment groups, given that a protective treatment effect will keep frailer individuals alive in the experimental group compared with the placebo/comparator group. Without adjustment, this will be reflected in hazard or HR estimates and conditioning on prognostic factors will change the value and interpretation of the HR.[18] This is known as noncollapsibility: a well-documented concept in causal literature that is less commonly considered for survival extrapolation. Because treatment effect waning happens at an individual level, analyses exploring it should be based on conditional, rather than marginal, HRs.

If individual-level treatment effects disappear completely, having frailer individuals in the surviving treatment group means the unadjusted (strictly speaking, marginal) HR will be >1. Constraining the marginal HR to 1 will not be equivalent to loss-of-treatment effect within each participant. In a cost-effectiveness setting, this means estimates of "survival difference under individual-level treatment effect waning" can be less conservative than intended, overstating long-term treatment effects and underestimating incremental cost-effectiveness ratios (ICERs). These issues are formalized below and explored through simulation to demonstrate behavior of marginal HRs under true conditional (individual-level) HR waning and the potential biases induced by constraints.

Although, for simplicity, the following technical details and proposed simulation study are described using survival terminology, points made apply irrespective of the negative event (eg, disease progression) modeled.

Methods

Hazard Rates

The marginal hazard rate at time t gives the population-level instantaneous failure rate of those still at risk (ie, requiring survival time $>t$). Conditioning on covariates returns the hazard rate for a subset of the population with given covariate values. The latter is estimated through modeling or stratification. The former can be obtained from an unadjusted analysis or using covariate-adjusted marginal treatment effect estimators, covered by Morris et al.[19]

A hazard rate may be conditional on some covariates and hence marginal over all others (either measured/known or unmeasured/unknown). It will never be conditional on all predictors of survival in a real-world context — for example, data on predictors such as full genome and entire dietary history are likely to be impossible to include. This means that a conditional hazard rate might not necessarily correspond to an individual participant, but a subgroup. For simplicity, this is still referred to as an individual-level hazard rate here.

In an RCT, hazards will be conditioned on treatment assignment (X) for contrasts. Conditional versus marginal hazards in this setting refer to conditioning on other prognostic covariates (Z), see Eq. (1) versus (2), respectively.

$$h(t|X=x,Z=z)=\lim_{\delta t\to 0}\frac{P(t\le T<t+\delta t \mid T\ge t,\ X=x,Z=z)}{\delta t} \quad (1)$$

$$h_m(t|X=x)=\lim_{\delta t\to 0}\frac{P(t\le T<t+\delta t \mid T\ge t,\ X=x)}{\delta t} \quad (2)$$

Selection Bias and Noncollapsibility

The condition on survival time $T \ge t$, as in Eq. (1) and (2), can induce imbalanced comparisons for $t > 0$, even with perfect randomisation at baseline. The overall "surviving" sample is selected for healthier participants as the frailest, eg, oldest or with more severe disease, die sooner. Assuming presence of a protective treatment effect reducing treatment-group mortality rates, this will happen faster on the placebo arm than the treatment arm. As such, the distribution of baseline prognostic factors in treatment arm survivors for $t > 0$ becomes unhealthier than the control group. This is referred to as HR selection bias.[17]

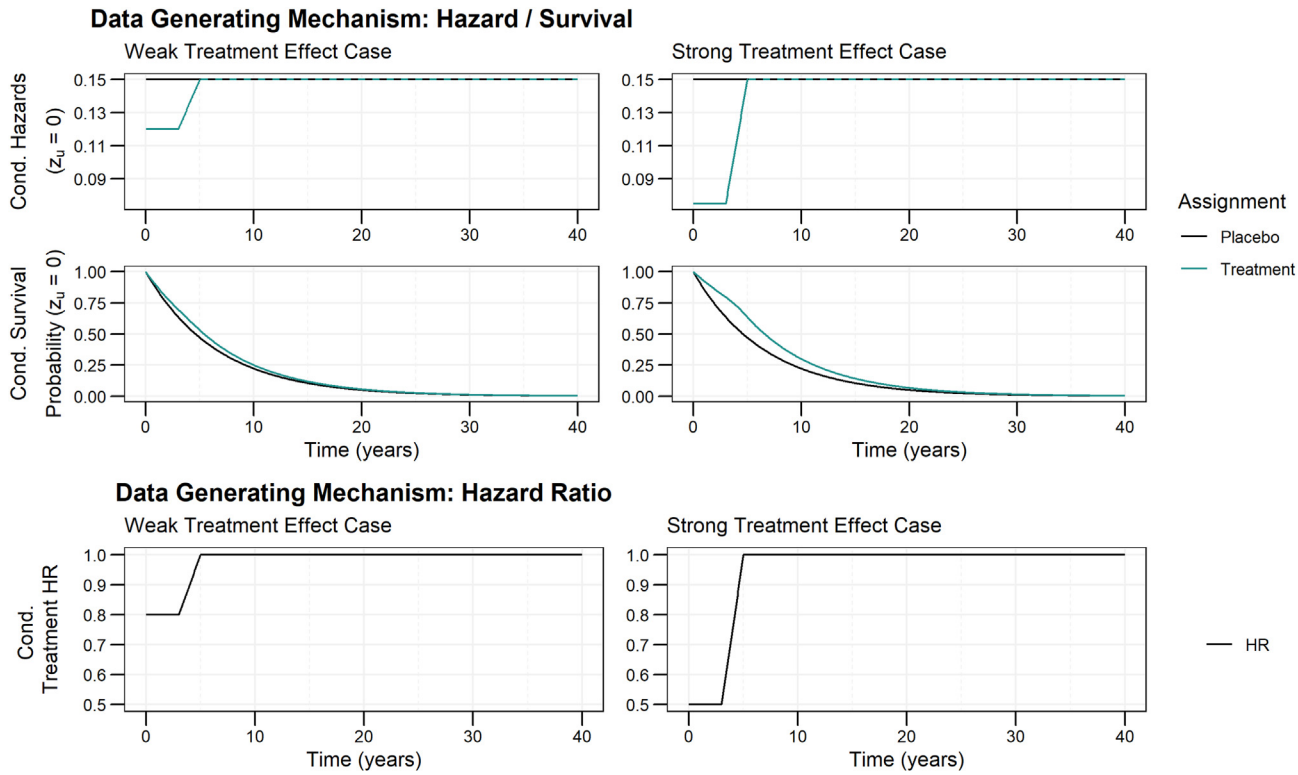
In these circumstances, marginal hazards or HRs over time would reflect these systematic differences in post-baseline survivor prognostic factor distributions. This leads to hazard and HR noncollapsibility, whereby conditioning a hazard on any prognostic factor changes the value and interpretation of HR, even if the factor is unrelated to treatment assignment. This is introduced conceptually here; refer to references[18,20] for a complete discussion. Noncollapsibility also holds for odds ratios but not for hazard differences (assuming continuous time) or risk differences or ratios.[21]

For illustration, consider a constant, conditional (individual-level) treatment effect and a single frailty "score" for each participant summarizing their overall hazard of death, excluding treatment assignment (a combination of age, sex, disease severity, etc.), with a lower score representing a lower risk of death. As time progresses, the average score in survivors in the comparator group will drop faster than in the treatment group, becoming less frail on average; post-randomization comparisons will suffer from selection bias. Hazard rates stratified by treatment but marginal over (eg, unadjusted for) this score will reflect this differential change in frailty over time and marginal hazards will drop faster in the control group than the treatment group. This induces non-PH (proportional hazards) in the marginal HR, with the marginal HR equal to the conditional HR at $t = 0$ but with the marginal treatment effect attenuated (ie, HR closer to 1) for all $t > 0$, and leads to HR noncollapsibility. If PH were assumed, the marginal HR estimate would be closer to 1 than the conditional value; see Appendix Figure 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.12.008. It is important to note the potential for this behavior to be mistaken for treatment effect waning, even under a time-invariant individual-level treatment effect.

For a simple, real-world example of this, consider the trial of levamisole and fluorouracil (Lev+5FU) versus placebo in colon cancer patients, reported by Laurie et al.[22] At baseline, mean ages were balanced between treatment groups (59.5 vs 59.9 years). Lev+5FU reduced the mortality rate and the surviving sample became selected for healthier participants at different speeds in each treatment group: 8 years after randomization, the mean baseline age was 58.9 years in the Lev+5FU group and 55.4 years in the placebo group.

In the context of treatment effect waning, diminution of a fully conditional HR corresponds to the intervention losing effect at the individual-level, across all covariate profiles. In this case, a marginal HR may exceed 1 because of covariate imbalances between treatment groups that occur over time, as previously discussed. Waning of a marginal HR is a combination of changing conditional treatment effect and prognostic factor distributions. In the presence of the

Figure 1. True conditional (on $z_u = 0$) hazard, survival probabilities and treatment-specific hazard ratio for varying treatment effect size.



Cond. indicates conditional; HR, hazard ratio.

expected distribution imbalances, this implies retention of some individual-level treatment effect, offsetting distribution differences. Hence, constraining the marginal HR to 1 is not equivalent to a scenario where there is no individual-level treatment benefit after a specified time point, and such analyses do not directly address the scenario recommended by NICE: that of no further treatment benefit beyond treatment discontinuation. If, instead, the intention behind treatment effect waning analyses is simply to investigate the sensitivity of cost-effectiveness results to scenarios where the treatment effect reduces by some substantial, unspecified amount, constraining the marginal HR to 1 may still be informative. However, for high-quality evidence-based decision making, it seems reasonable to expect more accurate and more precisely interpretable analyses. Exploring a scenario where the individual-level treatment effect disappears at a specified time point is much more precisely interpretable than a scenario where the marginal HR is constrained to 1.

Effect modification of the conditional HR by a prognostic factor could affect these arguments in various ways. The systematic differences in selection effects between treatment groups may be weakened, exacerbated, or potentially reversed entirely, depending on the type of effect modification present. The presence of effect modification in RCTs complicate interpretation of estimates far beyond implications for the work presented here; this is covered in further detail in Discussion section.

It is possible to obtain marginal hazard rates from a model incorporating covariates Z using regression standardization (covariate-adjusted marginal treatment effect estimation).[19] This is important when conditioning is required but a marginal (population-level) estimate is of interest. This estimate, Eq. (3), is an average of the conditional hazards, weighted by the probability of an individual with that covariate pattern still being at risk to contribute to the marginal hazard. The equivalent estimator of marginal survival probability is defined simply as the average of conditional survival probabilities.

$$\widehat{h}_M(t|X=x) = \frac{\frac{1}{N}\sum_{i=1}^{N} S(t|X=x, \boldsymbol{Z}=\boldsymbol{z_i}) h(t|X=x, \boldsymbol{Z}=\boldsymbol{z_i})}{\frac{1}{N}\sum_{i=1}^{N} S(t|X=x, \boldsymbol{Z}=\boldsymbol{z_i})} \quad (3)$$

A simulation study is proposed to describe behavior of marginal HRs under conditional efficacy waning and to evaluate the impact of forcing a marginal, instead of conditional, HR to 1.

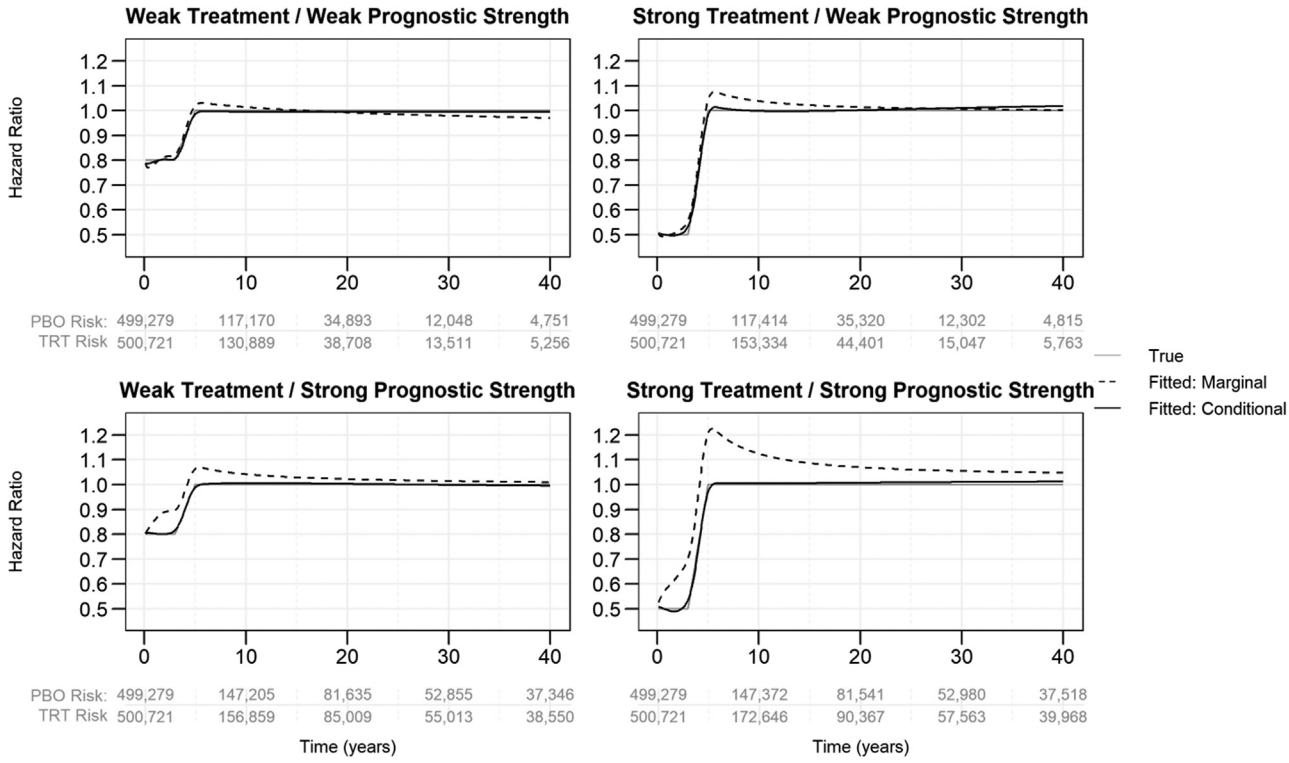
Simulation Study

This simulation study is presented using the Aims, Data-generating mechanisms, Estimands, Methods, Performance measures (ADEMP) framework.[23] Marginal herein refers to a HR conditioned only on treatment assignment while conditional assumes all other predictors of survival can be measured and are adjusted for.

Aims

Aim 1 is to demonstrate behavior of the marginal HR under scenarios where the conditional HR wanes. Aim 2 is to describe the bias in "RMST (restricted mean survival time) difference under loss of individual-level treatment effect" induced by forcing the marginal HR to 1. This reflects the case in which treatment effect waning constraints, intending to evaluate loss of individual-level effect, are applied to marginal (rather than conditional) HRs.

Figure 2. Time-varying modeled conditional and marginal hazard ratios for 4 scenarios, 40-year censored data, compared with the true conditional HR used for simulation. Baseline knots evenly spaced between minimum and maximum event times and treatment effect knots spaced evenly between minimum event time and year 6 (including a knot at 6 years) with no further knots included between 6 years and the upper boundary knot (maximum event time <40).



PBO indicates placebo; TRT, treatment. True values are largely obscured by modeled conditional estimates.

Data-generating mechanisms

1 000 000 (n_{obs}) participants were generated with a random binomial (p = 0.5) treatment assignment (X = 0/1 for placebo/treatment) and a random standard normal heterogeneity measure, z_u. From these covariate values, survival times were generated (using techniques outlined in reference[24,25]) based on (1) an exponential distribution (λ = 0.15), implying constant individual-level baseline (placebo) hazard, (2) a constant, protective treatment effect from 0 to 3 years, linearly approaching no-effect (HR = 1) between 3 to 5 years, maintained for the rest of follow-up, and (3) a multiplicative impact of heterogeneity on hazards: $h(t) = \lambda\ HR_X(t)\ exp(\beta_z z_u)$, in which $HR_X(t) = 1\ \forall t$ for placebo participants and represents the protective effect and subsequent waning structure for treatment participants. Two effect sizes were used for the conditional HR pre-waning: HR 0.8 or 0.5 for weak and strong treatment effects, respectively, and 2 levels of heterogeneity were chosen: β_z = 0.4 or 1.2 for weak and strong prognostic strengths, respectively. This led to 4 scenarios for comparison. Refer to Figure 1 for hazard and HR structures used (for mean value of z_u) and Appendix Figure 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.12.008, showing the impact of heterogeneity. A single large sample (n_{sim} = 1) was utilized because the aim was to demonstrate the bias associated with using a marginal estimate to assess individual-level waning, rather than investigating issues such as coverage or relative precision. Although not the most common approach used in simulation studies, this aligns with standards for assessing large-sample bias.[23] Appendix Figure 3, in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.12.008, includes a justification of these heterogeneity effect sizes using a real data set.[26]

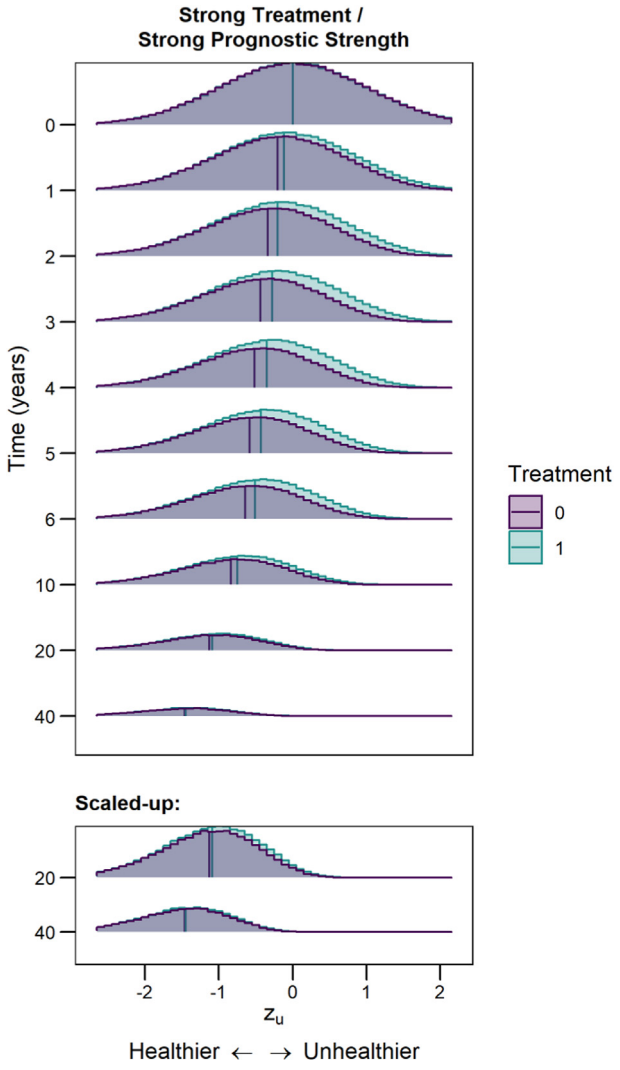
In practice, survival estimates beyond the RCT time period would be based on extrapolations; however, the aim of this work was not to demonstrate complexities with extrapolation techniques but with the assumptions made for extrapolation.

Aim 1: Behavior of the marginal HR under a waning conditional treatment effect

Estimands. The estimands for aim 1 were time-varying 40-year conditional and marginal HRs under each scenario.

Methods. Log-hazard-scale FPMs (flexible parametric models),[27] fitted to simulated data with all event times over 40 years censored, were used to estimate time-varying HRs. The 40-year cutoff approximates a lifetime horizon given the selected parameters. FPMs involve the use of natural cubic splines of log-time to model baseline hazards and time-varying treatment effects, here using 2 and 5 internal knots, respectively, and allow flexibility in the modeling of hazard structures. They were used here such that time-varying HRs could be accurately modeled, permitting an accurate exposition of the conditional and marginal HRs in simulated scenarios. Knots were chosen to achieve the maximum flexibility that did not show signs of overfitting. Because survival estimates (on full data) are known to be robust to hazard knot specification,[28] this decision is not expected to have undue influence on results.

Figure 3. Ridgeline plots of survivor z_u distributions, mean value identified, 40-year censored data.



0 indicates placebo; 1, treatment.

Aim 2: Impact of constraining the marginal HR to 1 when assessing conditional treatment effect waning

Estimands. For aim 2, the estimand was the marginal (restricted mean survival time difference [ΔRMST]: treatment − placebo) under loss of individual-level treatment effect.

Methods. True marginal, treatment-specific RMST values were calculated via numerical integration (between 0 and 40 years) of treatment-stratified Kaplan-Meier[29] curves fitted to each of the 4 scenario data sets.

To constrain marginal HRs to 1, marginal hazards in the treatment group, estimated in aim 1, were replaced by hazards estimated for the placebo group for all $t > 5$. Marginal RMSTs were derived using bounded (0 to 40) numerical integration of survival curves calculated numerically from the altered marginal hazards.

Performance measures. Bias in ΔRMST induced by marginal constraints under each scenario ("constrained marginal HR" ΔRMST − true ΔRMST, where true ΔRMST was calculated under simulated attenuation of the conditional treatment effect).

All analyses were carried out in R Statistical Software (v4.1.2).[30] Data were simulated using simsurv[31] and FPMs fitted using survPen.[32]

Results

Aim 1: Behavior of the Marginal HR Under a Waning Conditional Treatment Effect

Figure 2 shows the 40-year true conditional and estimated marginal and conditional HRs for the 4 parameter combinations. The estimated marginal HR approaches 1 over the first 3 years, despite constant conditional HRs, before increasing more rapidly as the conditional treatment effect wears off. It takes a value >1 from around the point that the true conditional HR reaches unity. After peaking, it then becomes closer to the conditional estimate with increasing follow-up time. This reflects changes in both conditional hazards and prognostic distributions. The amount by which the marginal treatment effect is attenuated (or equivalently the HR increases toward 1) from 0 to 3 years and the value the marginal HR reaches following this point is highly dependent on the initial conditional treatment effect size and prognostic strength of z_u, with maximum value just over 1.2.

Figure 3 shows empirical distributions of z_u values in surviving participants from the strong treatment, strong prognostic effect scenario. Over 0 to 3 years, the mean z_u value of survivors in the placebo group approaches healthier, more negative values faster than in the treatment group, despite near perfect balance at baseline. After 3 years, the rate at which the mean z_u value in treatment-group survivors approaches healthier values accelerates, and by the 40-year time point mean values of z_u are almost identical between survivors in each treatment group.

Aim 2: Impact of Constraining the Marginal HR to 1 When Assessing Conditional Treatment Effect Waning

Figure 4 shows the unconstrained marginal HRs versus those that are constrained to 1 from 5 years for the 4 scenarios. When a marginal HR is constrained to 1, the HR is reduced, removing the expected marginal HR estimates exceeding 1 and constraining marginal treatment-group hazards lower than their unconstrained values, see Figure 5.

Table 1 shows true versus constrained 40-year marginal RMST estimates for the 4 scenarios.

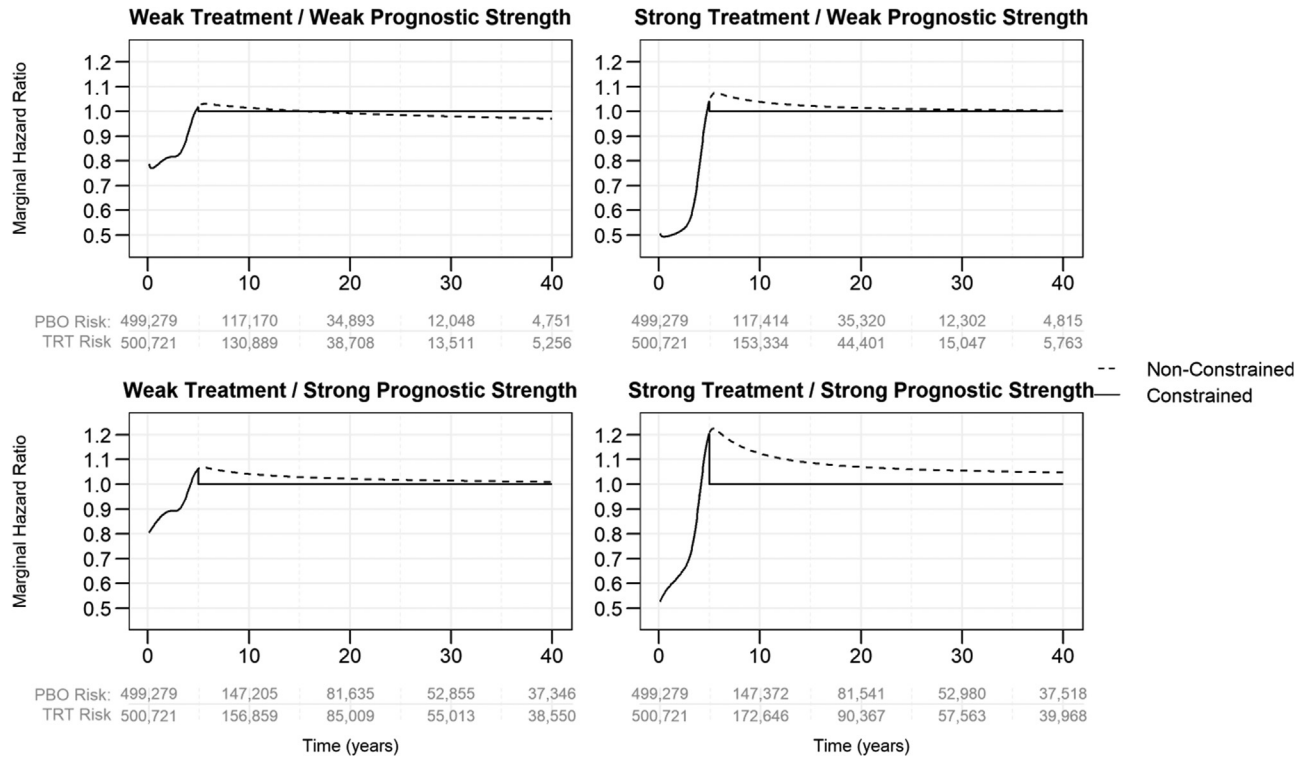
For the strong treatment and strong prognostic effect case, the true marginal ΔRMST was 1.4 years, whereas ΔRMST from the constrained model was 2.2 years. This constitutes a 40-year ΔRMST increase of 0.8 years (57%) due to artificial reduction in treatment-group hazards. Bias increased with initial treatment effect strength and prognostic strength of z_u, with the smallest ΔRMST bias observed in the weak treatment and weak prognostic effect scenario.

For completeness, bias in ΔRMST was calculated after constraining fully conditional estimates, matching the mechanism used for generating the data. As expected, this led to low bias (<0.04 years under all scenarios). Because of the practical impossibility of deriving an estimate conditioned on all prognostic covariables these results are not included in Table 1.

Discussion

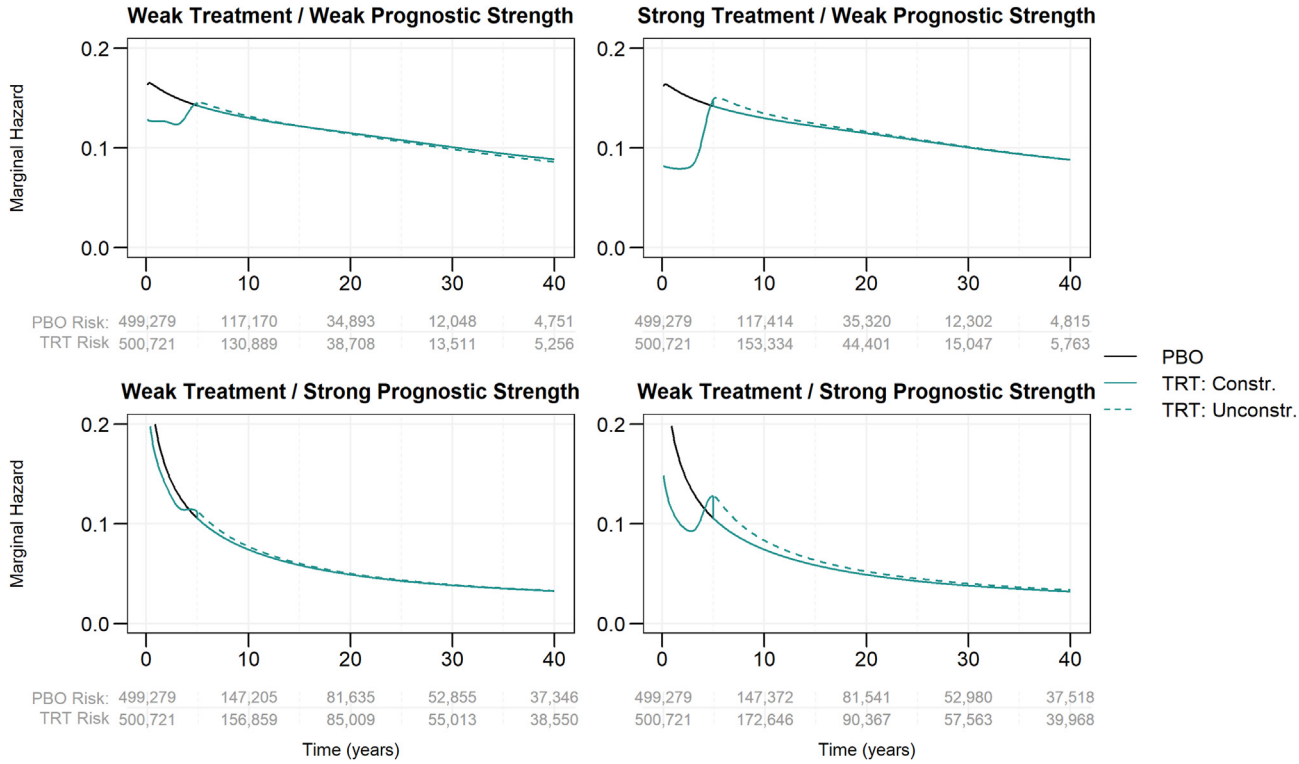
In this simulation exercise, a marginal HR was shown to take a value >1 from the point that a protective, individual-level treatment effect disappeared. This can be attributed to systematic

Figure 4. Time-varying non-constrained versus constrained marginal hazard ratios for 4 scenarios, 40-year censored data.



PBO indicates placebo; TRT, treatment.

Figure 5. Time-varying non-constrained versus constrained marginal hazards for 4 scenarios, 40-year censored data.



PBO indicates placebo; TRT, treatment; Constr., constrained; Unconstr., unconstrained.

Table 1. Marginal 40-year placebo and treatment (Δ, treatment − placebo) restricted mean survival time truths and estimates for the 4 scenarios and increase in ΔRMST attributable to constraining the HR to 1, 40-year censored data.

Treatment/prognostic effect strength scenario	40-year marginal PBO/TRT (Δ) RMST (years)		
	True RMST	Constrained HR RMST	ΔRMST Bias (%)
Weak/Weak	7.1/7.7 (0.6)	7.1/7.8 (0.7)	0.1 (17)
Weak/Strong	9.5/10.0 (0.5)	9.5/10.2 (0.7)	0.2 (40)
Strong/Weak	7.1/8.8 (1.7)	7.1/9.0 (1.9)	0.2 (12)
Strong/Strong	9.5/10.9 (1.4)	9.5/11.7 (2.2)	0.8 (57)

ΔRMST indicates treatment RMST-placebo RMST; HR, hazard ratio; PBO, placebo; RMST, restricted mean survival time; TRT, treatment.

differences in prognostic factor distributions in surviving participants over time. Over the first 3 years of a protective conditional treatment effect, survivors in the treatment group are selected for healthier covariate patterns at a slower rate than in the placebo group. Upon conditional treatment effect loss, the marginal hazards in the treatment group are greater than in the placebo group and the marginal HR is >1, reflecting the unhealthier covariate patterns. This higher mortality rate in the treatment arm means that, having diverged, the covariate distributions in survivors in each treatment group become more similar with increasing follow-up time and the marginal HR converges toward the conditional HR. Over time, the rate of convergence in covariate patterns slows; therefore, small differences and marginal HRs >1 can persist into long-term follow-up.

In real-life situations, data beyond trial follow-up will not be available. Thus, assumptions are required for extrapolation. Aiming to estimate the survival difference that would be observed under individual-level treatment effect waning, analysts may derive survival estimates from treatment-stratified (and otherwise marginal) hazards that are forced to be equal from a specific time point. However, this constitutes the artificial removal of post-baseline differences in patient characteristics between treatment arms that would cause, with true individual-level treatment effect waning, the unconstrained marginal HR to surpass 1.

When marginal HRs were constrained to 1, with marginal treatment-group hazards hence constrained to a value that was lower than the true marginal treatment-group hazard given a loss of individual-level treatment effect, survival difference (ΔRMST) was inflated by between 0.1 and 0.8 years (or by between 12 and 57%). This implies that if the loss of individual-level treatment effect is incorrectly modeled by setting a marginal HR to 1, the long-term survival benefits associated with the new treatment will be overestimated and ICERs will be underestimated, sometimes substantially.

The range in ΔRMST bias observed was due to varying prognostic strength of z_u and treatment effect strength over years 0 to 3. Appendix Figure 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.12.008 indicates that the strong prognostic effect case may be comparable with a disease with reasonable separation in participant prognosis (ie, by age or disease stage). Diminished biases for weak prognostic effect cases show that, with low prognostic strength in covariates over which the HR is marginal, setting the marginal HR to 1 may closely estimate effects under loss of individual treatment. The scope for bias associated with setting the marginal HR to 1 to approximate an individual-level loss-of-treatment effect is larger when strongly prognostic covariates exist. In practice, conditioning on important prognostic factors before enforcing treatment effect waning assumptions may more closely mimic the "weak prognostic effect of heterogeneity case" presented here, and bias in survival benefit estimates would be minimized.

We hence suggest that HRs used to model treatment effect waning are conditioned on all possible prognostic factors and conditional survival estimates under efficacy waning calculated. The extent of adjustment possible will depend on clinical context, participant burden, and data collected in trials, but often the most important prognostic factors generating significant heterogeneity between participants would be collected routinely, eg, age, sex, disease severity, and factors used for randomization stratification.

As economic models used in HTA are used to inform population-level decision making, marginal estimates are generally preferable. Hence, we propose the use of regression standardization, introduced previously, to derive survival estimates stratified only by treatment. This conserves precision gains associated with covariate adjustment and allows marginal estimates to be used in the economic model. Full reporting of covariate adjustment applied to HRs being waned is important such that potential bias can be assessed.

Implementation of these recommendations ultimately falls on those with access to individual participant data). The best way for groups without individual participant data access to assess individual-level treatment effect waning for HTA, beyond just acknowledging the scope for bias, is not clear. A plausible way might be to constrain a marginal HR to several values between 1 and 1.2 and assess impacts; however, because the true marginal value is time-varying and highly dependent on many factors, this is by no means ideal and would require further research to confirm validity.

Our findings have important implications for HTA economic modeling, irrespective of the structure of the economic model implemented; whether in partitioned survival models or state-transition or multistate models, if treatment effect waning is modeled by setting marginal HRs to 1, bias will result. To assess the potential impact of these findings, all NICE HTAs published between June 2022 and June 2023 reporting waning were considered. We reviewed summary guidance documents for each appraisal and all available appraisal documents for the 5 most recent appraisals.[33-37] None made reference to covariate adjustment in economic models in summary guidance documents. In 1 of the 5 recent appraisals, waning was applied to a HR calculated using inverse probability of censoring weighting from an indirect comparison, whereas the rest appeared to wane fully marginal HRs.

Although survival and hence heterogeneity in life expectancy have been referred to throughout, this work also applies more widely to any efficacy waning applied to a HR. For an alternative

negative event, such as disease progression, heterogeneity would be redefined appropriately, eg, corresponding to between-patient heterogeneity in risk of progression. In multiple sclerosis, disability progression according to the Expanded Disability Status Scale may be the event, and factors prognostic of progression (from involvement of multiple systems to sex) would jointly define heterogeneity. In these cases, failure to condition HRs on sufficient prognostic factors before implementing treatment effect waning would also lead to estimates far from the true time-to-event (eg, progression) under loss of individual-level efficacy.

Our use of a simulation study with a large sample facilitates clear exposition of differences between conditional and marginal HRs and the impacts of applying constraints to marginal HRs when treatment effect waning works on an individual level. However, limitations are associated with this design. We considered a limited number of scenarios, and the biases reported are a function of the choice of absolute hazards, the duration, magnitude, and waning structure of protective effects, the heterogeneity modeled, and the way this was assumed to influence survival. For each parameter, values were selected to approximate real-life situations, with appropriateness of the prognostic strength of heterogeneity used in simulations demonstrated; however, alternative values may be similarly valid, limiting generalizability of these results to some extent. Utilizing a single sample of 1 million observations departs from real-life approximation and means that sampling effects could not be assessed. The increased variability in bias observed in samples of a more moderate size is not clear. Nonetheless, using a simulation study conferred advantages and allowed us to achieve the primary objective of this article — to demonstrate the potential for bias because of misspecification of estimates when implementing treatment effect waning scenarios to inform healthcare decision making.

An assumption in this simulation study is that of no treatment effect modification of the conditional HR. With effect modification, the selection effect and hence the amount by which the marginal HR exceeds 1 under loss of individual-level treatment effect could be exacerbated or attenuated (or potentially reversed, with marginal HRs < 1), depending on which participants benefit most from treatment. The existence of treatment effect modification in RCTs introduces additional challenges to healthcare decision making than the treatment effect waning scenarios focused on in this article. However, even under treatment effect modification, biases in estimates of "survival difference under individual loss-of-treatment effect" are still highly likely if the marginal HR is constrained to 1.

There are a variety of possible effect modification scenarios and investigation of these is beyond the scope of this work. Additionally, the wider discussion on circumstances when covariate adjustment is appropriate in RCT analyses and HTA, requiring modeling choices for conditional relationships (eg, interactions with continuous covariates or nonlinear effects), is beyond the scope of this article. Thus, the discussion here was limited to how to appropriately model individual-level treatment effect waning in an RCT setting.

In summary, treatment effect waning can garner a great deal of discussion in HTA. There is often little evidence as to whether waning truly occurs across different disease areas, much less on appropriate time periods for modeling. However, scenarios including treatment effect waning are commonly considered in HTAs, and a common approach used has been shown to induce bias. In HTA, discussion primarily focuses on if, when, and how (ie, instantaneously or gradually) treatment effect waning happens[15]; these are important points, but we highlight another issue related to efficacy waning that has so far been overlooked.

We reiterate that a fully conditional, individual-level HR for constraint is not achievable in practice and that the bias induced by anything less is impossible to quantify in a real-life scenario. However, it is possible to minimize this bias through the adjustment of HRs used in waning analyses. What actually constitutes a real-life case where unadjusted scovariates are such that bias is negligible is less clear, depending on clinical context and level of heterogeneity in the patient population (and how much of this can be modeled).

Conclusions

Important differences exist between extrapolations that assume marginal versus conditional treatment effect waning. When a marginal HR is constrained to 1 to assess efficacy under individual-level treatment effect waning, the survival benefits associated with the new treatment (assuming it prolongs life) will be overestimated and ICERs will be underestimated. We propose HRs are adjusted for all possible prognostic covariates before use in treatment effect waning scenarios, with regression standardization used to return to marginal estimates for use within economic models to inform healthcare decision making.

Author Disclosures

Links to the disclosure forms provided by the authors are available here.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2023.12.008.

Article and Author Information

Accepted for Publication: December 15, 2023

Published Online: January 20, 2024

doi: https://doi.org/10.1016/j.jval.2023.12.008

Author Affiliations: Biostatistics Research Group, Department of Population Health Sciences, University of Leicester, Leicester, England, UK (Jennings, Rutherford, Sweeting, Lambert); School of Health and Related Research, University of Sheffield, Sheffield, England, United Kingdom; Delta Hat Limited, Nottingham, England, UK (Latimer); Statistical Innovation, AstraZeneca, London, England, UK (Sweeting); Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden (Lambert).

Correspondence: Angus C. Jennings, MSc, Department of Population Health Sciences, University of Leicester, University Road, Leicester, LE1 7RH, England, United Kingdom. Email: acj23@leicester.ac.uk

Author Contributions: *Concept and design:* Jennings, Rutherford, Lambert
Acquisition of data: Lambert
Analysis and interpretation of data: Jennings, Rutherford, Latimer, Sweeting, Lambert
Drafting of the manuscript: Jennings, Latimer, Rutherford, Sweeting, Lambert
Critical revision of paper for important intellectual content: Jennings, Rutherford, Latimer, Sweeting, Lambert
Statistical analysis: Jennings, Rutherford, Lambert
Obtaining funding: Jennings, Rutherford, Lambert
Supervision: Rutherford, Lambert

Funding/Support: This study was supported/funded by the National Institute for Health and Care Research (NIHR; AJ grant research number NIHR302030), Applied Research Collaboration East Midlands (ARC EM) and Leicester NIHR Biomedical Research Centre (BRC). NRL is funded by Yorkshire Cancer Research (Award reference number S406NL). The views expressed are those of the author(s) and not necessarily those of the NIHR, the Department of Health and Social Care, or Yorkshire Cancer Research. MS is a full-time employee of AstraZeneca.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of this work; generation, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Guyot P, Welton NJ, Ouwens MJ, Ades A. Survival time outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. *Value Health*. 2011;14(5):640–646.
2. NICE health technology evaluations: the manual. National Institute for Health and Care Excellence. https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation. Accessed June 22, 2023.
3. Kongnakorn T, Sarri G, Freitag A, et al. Modeling challenges in cost-effectiveness analysis of first-line immuno-oncology therapies in non-small cell lung cancer: a systematic literature review. *Pharmacoeconomics*. 2022;40(2):183–201.
4. Armoiry X, Wang-Steverding X, Connock M, et al. Is the assumption of waning of treatment effect applied consistently across NICE technology appraisals? A case-study focusing on disease-modifying therapies for treatment of multiple sclerosis. *Int J Technol Assess Health Care*. 2022;38(1):e83.
5. Willem L, Blommaert A, Hanquet G, et al. Economic evaluation of pneumococcal vaccines for adults aged over 50 years in Belgium. *Hum Vaccin Immunother*. 2018;14(5):1218–1229.
6. National Institute for Health and Care Excellence. *Ocrelizumab for treating primary progressive multiple sclerosis*. Technology Appraisal Guidance TA585; 2019. https://www.nice.org.uk/guidance/ta585. Accessed June 7, 2023.
7. National Institute for Health and Care Excellence. *Elosulfase alfa for treating mucopolysaccharidosis type 4A*. Highly Specialised Technologies Guidance HST19; 2022. https://www.nice.org.uk/guidance/hst19. Accessed June 7, 2023.
8. Toumi M, Dabbous O, Aballéa S, et al. Recommendations for economic evaluations of cell and gene therapies: a systematic literature review with critical appraisal. *Expert Rev Pharmacoecon Outcomes Res*. 2023;23(5):483–497.
9. Al Hamarneh YN, Johnston K, Marra CA, Tsuyuki RT. Pharmacist prescribing and care improves cardiovascular risk, but is it cost-effective? A cost-effectiveness analysis of the RxEACH study. *Can Pharm J*. 2019;152(4):257–266.
10. National Institute for Health and Care Excellence. Voclosporin with mycophenolate mofetil for treating lupus nephritis. *Technol Appraisal Guid Ta*; 2023:882. https://www.nice.org.uk/guidance/ta882. Accessed June 7, 2023.
11. National Institute for Health and Care Excellence. Finerenone for treating chronic kidney disease in type 2 diabetes. *Technol Appraisal Guid Ta*; 2023:877. https://www.nice.org.uk/guidance/ta877. Accessed June 7, 2023.
12. Willis M, Nilsson A, Kellerborg K, et al. Cost-effectiveness of canagliflozin added to standard of care for treating diabetic kidney disease (DKD) in patients with type 2 diabetes mellitus (T2DM) in England: estimates using the CREDEM-DKD model. *Diabetes Ther*. 2021;12(1):313–328.
13. National Institute for Health and Care Excellence. Abrocitinib, tralokinumab or upadacitinib for treating moderate to severe atopic dermatitis. *Technol Appraisal Guid Ta*; 2022:814. https://www.nice.org.uk/guidance/ta814. Accessed June 7, 2023.
14. National Institute for Health and Care Excellence. Difelikefalin for treating pruritus in people having haemodialysis. *Technol Appraisal Guid Ta*; 2023:890. https://www.nice.org.uk/guidance/ta890. Accessed June 7, 2023.
15. Kamgar F, Ho S, Hawe E, Brodtkorb T. EE228 A review of treatment effect waning methods for immuno-oncology therapies in National Institute for Health and Care Excellence technology appraisals. *Value Health*. 2022;25(12):S98.
16. Wiyani A, Badgujar L, Khurana V, Adlard N. How have economic evaluations in relapsing multiple sclerosis evolved over time? A systematic literature review. *Neurol Ther*. 2021;10(2):557–583.
17. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21(1):13.
18. Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom J*. 2021;63(3):528–557.
19. Morris TP, Walker AS, Williamson EJ, White IR. Planning a method for covariate adjustment in individually randomised trials: a practical guide. *Trials*. 2022;23(1):328.
20. Didelez V, Stensrud MJ. On the logic of collapsibility for causal effect measures. *Biom J*. 2022;64(2):235–242.
21. Sjölander A, Dahlqwist E, Zetterqvist J. A note on the noncollapsibility of rate differences and rate ratios. *Epidemiology*. 2016;27(3):356–359.
22. Laurie JA, Moertel CG, Fleming TR, et al. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. *J Clin Oncol*. 1989;7(10):1447–1456.
23. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–2102.
24. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24(11):1713–1723.
25. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*. 2013;32(23):4118–4134.
26. Royston P, Lambert PC. *Flexible parametric survival analysis using Stata: beyond the Cox model*. TX: Stata Press College Station; 2011:347.
27. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21(15):2175–2197.
28. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *J Stat Comput Simul*. 2015;85(4):777–793.
29. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457–481.
30. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021.
31. Brilleman SL, Wolfe R, Moreno-Betancur M, Crowther MJ. Simulating survival data using the simsurv R package. *J Stat Softw*. 2021;97(3):1–27.
32. Fauvernier M, Remontet L, Uhry Z, Bossard N, Roche L. survPen: an R package for hazard and excess hazard modelling with multidimensional penalized splines. *J Open Source Softw*. 2019;4(40):1434.
33. National Institute for Health and Care Excellence. Pembrolizumab with lenvatinib for previously treated advanced or recurrent endometrial cancer. *Technol Appraisal Guid Ta*; 2023:904. https://www.nice.org.uk/guidance/ta904. Accessed June 7, 2023.
34. National Institute for Health and Care Excellence. Ibrutinib with venetoclax for untreated chronic lymphocytic leukaemia. *Technol Appraisal Guid Ta*; 2023:891. https://www.nice.org.uk/guidance/ta891. Accessed June 7, 2023.
35. National Institute for Health and Care Excellence. Pembrolizumab plus chemotherapy with or without bevacizumab for persistent, recurrent or metastatic cervical cancer. *Technol Appraisal Guid Ta*; 2023:885. https://www.nice.org.uk/guidance/ta885. Accessed June 7, 2023.
36. National Institute for Health and Care Excellence. Ixazomib with lenalidomide and dexamethasone for treating relapsed or refractory multiple myeloma. *Technol Appraisal Guid Ta*; 2023:870. https://www.nice.org.uk/guidance/ta870. Accessed June 7, 2023.
37. National Institute for Health and Care Excellence. Nivolumab with fluoropyrimidine- and platinum-based chemotherapy for untreated unresectable advanced, recurrent, or metastatic oesophageal squamous cell carcinoma. *Technol Appraisal Guid Ta*; 2023:865. https://www.nice.org.uk/guidance/ta865. Accessed June 7, 2023.