Adjusting for Treatment Switching in Oncology Trials: A Systematic Review and Recommendations for Reporting

Sullivan TR, Latimer NR, Gray J, Sorich MJ, Salter AB, et al., Value in Health, 2020

ABSTRACT

Background: Treatment switching is a common occurrence in oncology trials. Although methods such as the rank preserving structural failure time model (RPSFTM) and inverse probability of censoring weights (IPCW) have been developed to address treatment switching, the approaches are not widely accepted within health technology assessment. This limited acceptance may partly be a consequence of poor reporting on their application. **Objectives:** To systematically review the quality of reporting on the application of the RPSFTM and IPCW approaches in published trials and industry submissions to The National Institute for Health and Care Excellence. **Methods:** Published trials and industry submissions were obtained from searches of PubMed and nice.org.uk, respectively. The quality of reporting in these studies was judged against a checklist of reporting recommendations, developed by the authors based on detailed considerations of the methods. **Results:** Thirteen published trials and eight submissions to nice.org.uk satisfied inclusion criteria. The quality of reporting around the implementation of the RPSFTM and IPCW methods was generally poor. Few studies stated whether the adjustment approach was pre-specified, over a third failed to provide any justification for the chosen method, and nearly half neglected to perform sensitivity analyses. Further, it was often unclear how the RPSFTM and IPCW methods were implemented. **Conclusions:** Inadequate reporting on the application of the RPSFTM and IPCW methods increases uncertainty around results, which may contribute to the limited acceptance of these methods by decision makers. The proposed reporting recommendations aim to support the improved interpretation of analyses undertaken to adjust for treatment switching.

INTRODUCTION

Treatment switching in randomized controlled trials (RCTs) occurs when patients discontinue their randomly assigned treatment and commence an alternative treatment. It is especially common in oncology trials, where control group patients are often permitted to switch to the experimental treatment following disease progression. There are both ethical and practical reasons to allow treatment switching. Ethically, it may be inappropriate to deny control group patients the experimental treatment should interim analyses suggest it is superior to control (1). From a practical viewpoint, allowing for treatment switching can enhance trial recruitment, as patients may be more willing to consent if they know they will receive the experimental treatment at some point during the trial (1).

Unfortunately, treatment switching can introduce complexities in estimating treatment effects for longer-term outcomes, most notably overall survival (OS). Suppose an experimental treatment extends OS and that control group patients benefit from switching to the experimental treatment. In this case, the observed OS difference between the experimental and control arms would be smaller in magnitude than what would have been seen had switching not occurred. Whether this is problematic depends on the population parameter of interest. In health technology assessment (HTA), judgments around the cost-effectiveness of introducing experimental treatments into clinical practice typically rely on accurate OS comparisons with current standard care, where switching to the experimental treatment would not be possible (2-5). Hence for the purpose of HTA decision-making, it is often desirable to adjust OS estimates to reflect what would have been observed had control group patients not switched treatments. It is worth noting that treatment switching in the opposite direction, from the experimental to the control treatment, does not usually pose the same problem for HTA decision-making. Typically, no adjustment for treatment switching would be necessary provided the switches reflect what might occur with the introduction of the experimental treatment into clinical practice, for example patients ceasing the experimental treatment and commencing existing (control) treatments due to disease progression or toxicity.

A variety of statistical methods have been proposed to adjust for treatment switching in oncology trials, or equivalently, to estimate a switching-adjusted estimand. Simple methods include censoring patients at the time-point of the switch, excluding switching patients from the analysis altogether, or modelling treatment as a time-varying covariate. Although commonly used (1), these methods are prone to selection bias, since patients who switch treatments tend to have a different OS prognosis than patients who remain on the control treatment or were

originally randomized to the experimental arm (1, 6, 7). More rigorous approaches designed to account for this selection bias are available, with the rank preserving structural failure time model (RPSFTM) (8), inverse probability of censoring weights (IPCW) (9), and two-stage adjustment (10) methods among those currently considered most promising. Simulation studies have shown that these methods tend to produce more accurate estimates of the switching-adjusted estimand than simple adjustment methods or a standard intention to treat (ITT) analysis, but their performance can be compromised when underlying assumptions are violated (10-12).

Analyses using more rigorous adjustment methods are regularly submitted to HTA agencies, in some cases providing enough evidence to alter reimbursement decisions. In the case of sunitinib versus best supportive care for the treatment of gastrointestinal stromal tumours, for example, the incremental cost effectiveness ratio (ICER) per quality adjusted life year (QALY) changed from £90,500 in an ITT analysis to £31,800 with the RSPFTM; this led to sunitinib being recommended for reimbursement by The National Institute for Health and Care Excellence (NICE)(13). Similarly, in the case of pazopanib for the treatment of advanced renal cell carcinoma, NICE recommended pazopanib for reimbursement based on cost-effectiveness estimates derived from a RPSFTM (14). Interestingly, cost-effectiveness estimates for pazopanib differed substantially between adjustment methods, with the ICER/QALY versus interferon-α estimated to be £38,900 using the RPSFTM and £72,300 using IPCW. This illustrates just how important the choice of adjustment method and its appropriate implementation can be to reimbursement decisions.

Despite their potential, the RPSFTM, IPCW, and two-stage adjustment methods are not, at present, widely accepted by HTA agencies (15). As little guidance exists on what should be reported following a switching-adjusted analysis, it is possible this limited acceptance is a consequence of poor reporting around model implementation details and the plausibility of underlying assumptions. Importantly, there are many decisions involved in fitting the RPSFTM, IPCW, and two-stage adjustment methods, each of which could influence final treatment effect estimates. For example, there are several structural models and fitting algorithms available for the RPSFTM, such that different implementations of the method could conceivably lead to different qualitative conclusions. A clear description of how a model was implemented, including whether the chosen approach was pre-specified, would alleviate concerns over selective reporting and improve confidence in results. Similarly, if the assumptions of a switching-adjusted analysis are well justified, and if treatment effect estimates are shown to be similar across a range of sensitivity analyses, decision makers are more likely to accept their results (5, 16).

To evaluate the quality of reporting on the application of switching adjustment methods, we undertook a systematic review of published RCTs and industry submissions to NICE. We were specifically interested in the quality of reporting around the nature of treatment switching in the trial, how the switching adjustment method was implemented, and whether model assumptions were justified and interrogated using sensitivity analyses. To assist the review process and provide guidance for future use of the methods, we also developed a checklist of recommendations for the reporting of switching-adjusted analyses.

METHODS

To give context to the aspects of reporting considered in the systematic review, we first describe the RPSFTM and IPCW methods, drawing attention to the model-fitting steps involved in their application. As two-stage adjustment has only recently been proposed in the statistical literature (10), we do not consider its use in the review.

Rank preserving structural failure time model

The RPSFTM uses a potential outcomes framework to estimate OS times that would have been observed had treatment switching not occurred. Suppose for the i^{th} randomized patient that their observed OS time T_i can be partitioned into the time spent on the experimental treatment T_{Ei} and the time spent on the control treatment T_{Ci}. In the simple one-parameter version of the RPSFTM, the potentially counterfactual survival time U_i that would have been observed had the i^{th} patient received only the control treatment is determined according to the structural model

$$U_i = T_{Ci} + T_{Ei}/AF \qquad (1)$$

where AF represents the acceleration factor due to the experimental treatment; the amount by which the experimental treatment expands or contracts survival times on the time scale. The model assumes that the AF due to the experimental treatment is constant over time for all patients no matter when it was first received, known as the "common treatment effect" assumption (1), with the effect applying immediately upon commencement and

ceasing immediately upon discontinuation of treatment. To illustrate how this works, suppose a common treatment effect holds and that the experimental treatment doubles the length of survival relative to control (i.e. AF = 2). Imagine also that a given patient receives the control treatment for one year ($T_C = 1$) and then switches to the experimental treatment and survives a further three years ($T_E = 3$), thus recording an OS time of four years. Counterfactually, had this patient not switched to the experimental treatment, under the structural model in (1) they would have instead survived $U = 1 + 3/2 = 2.5$ years.

In a randomized trial U should be equivalently distributed between randomized arms, and so the AF can be estimated by searching through a range of plausible values and choosing the value that produces the most statistically similar untreated survival times between randomized groups, a process known as g-estimation. Once estimated, the AF is used to calculate counterfactual survival times in switching patients, which are then included in a final outcomes (FO) model to produce a switching-adjusted treatment effect estimate. To account for the uncertainty in the estimated AF, confidence limits around the switching-adjusted treatment effect estimate are calculated by retaining the p-value from a standard ITT analysis or applying bootstrapping methods (17).

The RPSFTM can be implemented in many different ways. Importantly, several different structural models involving different assumptions are possible, for example models that allow for lagged effects of treatment or effects that differ in magnitude across randomized groups. In addition to the "as treated" structural model described above, the RPSFTM can also be applied on an "ever treated" basis, where patients are considered to remain on the experimental treatment irrespective of later treatment discontinuation (i.e. T_{Ei} in (1) is taken to indicate the time following the commencement of the experimental treatment). Such a model might be appropriate when the effect of the experimental treatment is expected to persist beyond its discontinuation. The RPSFTM can also be modified to allow the AF to differ between treatment switchers and those randomized to the experimental arm by some pre-specified amount. This type of model can be used to explore the sensitivity of results to the common treatment effect assumption, or as a primary method of analysis when the common treatment effect assumption is deemed inappropriate.

There are also a variety of g-estimation options available for the RPSFTM. The test statistic used to demonstrate equivalence in U across randomized arms could be obtained from a log-rank test, a Wilcoxon test, or a Wald test from an accelerated failure time or Cox proportional hazards model, with or without adjustment for covariates.

Further, different grid search algorithms are available for estimating the AF, for example searching in fixed steps or using interval bisection (18). Following estimation of the AF, counterfactual survival times in treatment switchers may be obtained with or without re-censoring applied. Re-censoring refers to the earlier censoring of counterfactual survival times in order to avoid bias due to informative censoring (see (19) for further details), however this process can also associated with a loss of long term survival information. Recent simulation evidence suggests that the RPSFTM should be implemented both with and without re-censoring applied (19).

Like any statistical analysis, the validity of the RPSFTM hinges on estimation performance and the suitability of underlying assumptions. G-estimation performance can be assessed by plotting potential values for the AF against the observed test statistic; if successful, the procedure should identify a unique solution where the test statistic equals zero. The success of g-estimation and the suitability of model assumptions can also be assessed by comparing counterfactual survival times between randomized groups using a Kaplan-Meier plot. Assuming randomization is successful in balancing prognostic variables, counterfactual survival times should be equivalently distributed across randomized groups. Given the untestable nature of the common treatment effect assumption, clinical input into its plausibility is also critical. If the beneficial effect of the experimental treatment is anticipated to be quite different between patients originally randomized to the experimental arm and patients who switch to the experimental treatment partway through the trial, then alternatives to the RPSFTM should be considered.

Inverse probability of censoring weights

Unlike the RPSFTM, which attempts to recreate the distribution of survival times had treatment switching not occurred, the IPCW method involves adjusting for the effects of switching during estimation of the treatment effect. In the context of treatment switching from the control to the experimental treatment, the IPCW method involves (a) censoring patients at the time of switching, and (b) addressing potential selection bias by reweighting remaining control group patients still at risk of death by the inverse of their probability of *not* switching. Higher weights are assigned to non-switching patients with 'similar' characteristics to switching patients, allowing these patients to represent both themselves and switching patients in the analysis (10). To satisfy an assumption of "no unmeasured confounders" (NUC), the weights should be calculated from a correctly specified model including all baseline and time-varying characteristics predictive of both treatment switching and OS; in general, this

necessitates extensive data collection. Another important requirement of the IPCW method is that the probability of treatment switching must always be less than one for all possible predictor combinations, otherwise weights cannot be estimated (20, 21).

Like the RPSFTM, there are numerous ways to implement the IPCW method. The first step in applying the method is to calculate time-varying weights for control group patients using a weight determining (WD) model (22). Options for this model include specifying time as continuous and fitting a Cox proportional hazards model for time to treatment switching, or working with discrete time intervals and using pooled logistic regression to model the odds of treatment switching within each interval. For pooled logistic regression, a choice must be made around the width of each discrete time interval and how to account for time in order to capture changes in the underlying hazard of treatment switching. A popular choice for accounting for time is to use a spline function, albeit this brings about additional choices concerning the complexity of the spline function and placement of knots (for recommendations on these choices, see (22)). After specifying the type of WD model, decisions must be made around which baseline and time-varying covariates to include, what functional form each should take, and how to address missing data on covariates. Once finalized, time-varying unstabilized weights are obtained from the WD model as the inverse of the probability of not switching. Optionally, these weights can be stabilized by multiplying them by the marginal probability of not switching treatments, as derived from a separate WD model containing only baseline predictors of switching (20). Following the estimation of weights (and assigning experimental arm patients a weight of 1), a FO model is fitted to the weighted data to produce a switching-adjusted treatment effect estimate. As with the WD model, either a Cox proportional hazards or pooled logistic regression model can be applied, with both approaches requiring adjustment for baseline predictors of treatment switching when stabilized weights are used. To account for uncertainty in the estimation of weights, p-values and confidence limits in the FO model can be calculated using either robust variance estimation or bootstrapping.

Evaluating estimation performance and the suitability of the NUC assumption is critical in judging the validity of the IPCW method. Estimation performance can be assessed by interrogating coefficient estimates and weights from the WD model, with implausible coefficient estimates or extreme weights indicative of an underlying problem with model specification. Extreme weights are a particular concern, as they suggest that the OS of just a few patients may be having an unduly large influence on the switching-adjusted treatment effect estimate. In addition to omitting predictors from the WD model or modifying its functional form, extreme weights might be

addressed by truncating their values at some upper limit; however, such an approach lacks theoretical justification and may introduce bias (23). Residuals from the WD model (e.g. martingale residuals) can also be useful in diagnosing issues with model specification. In relation to the suitability of the (untestable) NUC assumption, expert clinical opinion is once again key. Importantly, consideration should be given both to the covariates included in the WD model and how exactly treatment switching decisions were made in the relevant trial (1). It is worth noting that, in practice, satisfying the NUC assumption and maintaining estimation performance tend to be competing objectives, since it may not be statistically feasible to include all potential predictors in a WD model when there are few non-switchers. For this reason, other adjustment methods may be preferable in trials with high switching proportions (1, 10, 12).

Systematic review

To investigate the quality of reporting on the application of switching adjustment methods in practice, we undertook a systematic review of published RCTs and industry submissions to NICE. For the review of published RCTs, we included full-length articles where the RPSFTM or IPCW methods were applied to adjust for the effects of treatment switching on OS. For the review of NICE submissions, we considered technology appraisals (TAs) where the methods were applied to OS results in the 'Clinical Effectiveness' section of the initial submission. In submissions where switching adjustment methods were applied to multiple trials, data were extracted for the key trial, or for the first trial for which clinical effectiveness results were presented if there was no key trial. For both published RCTs and NICE submissions, studies were excluded if the relevant trial was not in oncology, if it was a pilot or dose-finding study, or where an adjustment method was used to account for switching from the experimental arm or to other non-randomized treatments. Specifically for the review of published RCTs, methodological papers involving a short example analysis of trial data were also excluded. No studies were excluded based on their publication date.

Searches of PubMed and nice.org.uk were conducted on the 22nd May 2018. PubMed search terms were based on the Cochrane sensitivity and precision maximising search strategy for randomized trials (24), with additional terms for "rank-preserving structural failure time" and "inverse probability of censoring" included; these additional terms were used exclusively for the nice.org.uk search. Following an assessment of eligibility, information from eligible studies was transcribed by a single reviewer (TRS) to a data extraction form developed

specifically for this review. Full texts were examined for the review of published RCTs, while for NICE submissions, data were extracted from the 'Clinical Effectiveness' section of the initial submission. Details reported in appendices (where available) were also included in the review process.

In designing the data extraction form for the systematic review, a checklist of reporting recommendations was developed. An initial set of recommendations was proposed by the first author (TRS) following a review of the methodological literature on treatment switching and reporting requirements for RCTs. The recommendations were discussed and revised on several occasions by the co-author team, with the group representing a mix of clinical trial statisticians, epidemiologists and health economists with experience reviewing and advising on the interpretation of RCTs with treatment switching to inform funding decisions. In finalizing the reporting recommendations, we attempted to address the following key aspects of a switching-adjusted analysis: the nature and extent of treatment switching, data available for adjustment, implementation of the chosen adjustment method, impact of adjustment, suitability of model assumptions and results of sensitivity analyses.

RESULTS

Reporting recommendations

In Table 1 we offer guidance on what should be reported following a switching-adjusted analysis. The list of recommendations includes items that apply to all switching-adjusted analyses and items specific to individual methods of adjustment. It is worth noting that several of our suggestions represent best practice for the analysis of randomised trials and are not unique to switching-adjusted analyses (25, 26). Explanation and elaboration for the reporting items are provided in Appendix A. Appendix B provides an overview of two-stage adjustment and offers recommendations for reporting on this method. In Appendix C, we demonstrate the application and utility of the reporting recommendations through the analysis of a case study.

<Insert Table 1 here>

Systematic review

The electronic search of PubMed identified 56 articles, of which 13 (27-39) were included in the review following an assessment of eligibility (Figure 1A). A total of 23 TAs were found in the search of nice.org.uk, with 8 of these (14, 40-46) satisfying eligibility criteria (Figure 1B). In total 16 RCTs were represented across the 21 included studies, with four trials presented in both a published article and a NICE TA, and one trial presented across two published articles.

<Insert Figure 1 here>

Key characteristics of the included trials are presented in Table 1. The median number of randomized participants was 416 for published articles and 326 for NICE TAs, with a median of 64% (range 23% to 87%) of control group patients switching treatments in both settings. Treatment switching was permitted following disease progression in the majority of published articles (69%) and NICE TAs (88%), with termination of the double-blind phase of the RCT the next most common reason for allowing treatment switching. Just one published article failed to describe the conditions under which control group patients were permitted to switch treatments. Most included studies presented treatment effect estimates from a standard ITT analysis for comparison (92% and 100% for published articles and NICE TAs, respectively). Conversely, few studies stated that the switching-adjusted approach was pre-specified, and in several cases the choice of adjustment approach was not justified at all (some justification was provided for 54% and 75% of published articles and NICE TAs, respectively). Finally, 46% of published articles and 63% of NICE TAs conveyed the impact of adjustment by visually comparing observed and switching-adjusted survival times in a Kaplan-Meier plot.

<Insert Table 2 here>

Among included studies, 9/13 published articles and 6/8 NICE TAs used the RPSFTM. Table 3 summarizes the quality of reporting around the implementation of the RPSFTM in these studies. As shown in the table, only 67% and 33% of published articles and NICE TAs, respectively, described the structural model assumed in the main analysis. Less than half stated the metric used to demonstrate equivalence during g-estimation, just one NICE TA (17%) described the grid-search algorithm, and not a single study plotted g-estimation results. Similarly, the estimated AF from g-estimation, which for the "as treated" structural model conveys the causal effect of the experimental treatment, was infrequently reported. It was unclear whether re-censoring had been applied in 44%

and 67% of published articles and NICE TAs, respectively, and only three published articles assessed the performance of g-estimation by comparing counterfactual survival times between groups. Retaining the ITT p-value was the most commonly reported method for calculating confidence intervals in the FO model, but again this aspect of the analysis was not always reported. Only two published articles (22%) and one NICE TA (17%) stated the baseline variables adjusted for in the FO model. Finally, no studies directly assessed the sensitivity of results to the common treatment effect assumption by allowing the AF to differ between switchers and those randomized to the experimental arm.

<Insert Table 3 here>

The IPCW method was used less frequently than the RPSFTM, with 6/13 published articles and 4/8 NICE TAs employing this approach. The quality of reporting around the implementation of IPCW is summarized in Table 4. Excluding four studies where the type of WD model was unclear, pooled logistic regression was a more common choice than the Cox proportional hazards model for calculating weights. Of the five studies employing pooled logistic regression, only two (one published article and one NICE TA) fully detailed the width of the discrete time interval and how time was accounted for in the model (one used a restricted cubic spline, the other linear and quadratic terms). Encouragingly, three of the four NICE TAs (75%) listed all the covariates that were considered in the development of WD models, described the frequency of measurements for time-varying covariates, and indicated that stabilized weights were used. Conversely, these characteristics were generally overlooked in the published articles. The distribution of weights and coefficient estimates from the WD model were poorly reported on, particularly for published articles. Lastly, scant details were provided on the FO model, with 70% of included studies failing to indicate the method for calculating confidence intervals around the estimated treatment effect and which baseline variables were adjusted for.

<Insert Table 4 here>

Seven published articles (54%) and five NICE TAs (63%) reported undertaking sensitivity analyses where an alternative statistical method was used to estimate the switching-adjusted estimand. Among studies that used the IPCW method, sensitivity analyses included modifying the specification of the WD model (one study), fitting the RPSFTM (four studies), and modelling treatment as a time-varying covariate (one study). Among studies using

the RPSFTM, methods of sensitivity analysis included varying the implementation of the RPSFTM (e.g. structural model or use of re-censoring; seven studies), fitting the IPCW method (four studies), two-stage adjustment (three studies), censoring patients at the time-point of the switch (three studies), and modelling treatment as a time-varying covariate (two studies).

DISCUSSION

In this article we reviewed the quality of reporting on the implementation of switching adjustment methods in oncology trials, both in the published literature and in industry submissions to NICE. With a median of 64% of control group patients switching treatments, inadequate handling of switching could have led to considerably biased estimates of the switching-adjusted estimand. This underscores the importance of appropriate adjustment for treatment switching in these trials, particularly in the context of HTA decision-making. Despite this, the quality of reporting around the implementation of the RPSFTM and IPCW methods was generally poor. Few studies stated whether the adjustment approach was pre-specified, over a third failed to justify the chosen method, and nearly half neglected to perform sensitivity analyses. Further, it was often unclear how the RPSFTM and IPCW methods were implemented, making it difficult to judge the validity of resulting treatment effect estimates. Overall, there is considerable scope for improvement.

Among studies applying the RPSFTM, it was especially concerning to find such a large proportion failing to describe the structural model assumed. Although conceivably the standard "as treated" structural model may have been assumed in all these studies, we believe the structural model should always be explicitly stated given its role in defining the assumptions of the analysis. Another major concern was the lack of reporting around model diagnostics, particularly in the AF resulting from g-estimation and in comparisons of counterfactual survival times between randomized groups. Such diagnostics give important insight into estimation performance and the validity of underlying assumptions. In studies applying the IPCW method, an alarming finding was the high proportion failing to provide coefficient estimates from the WD model. Since the validity of this method depends almost entirely on appropriate specification of the WD model, the magnitude and direction of coefficient estimates from this model should be examined for plausibility. A further concern was the lack of reporting on the distribution of weights, which, like coefficient estimates, can be helpful in identifying problems with the WD model.

Another important shortcoming identified in the review concerned the use of sensitivity analyses. In particular, sensitivity analyses were not routinely presented, and when they were, they often involved simple adjustment methods known to be susceptible to selection bias (1, 6, 7). We believe that both modifying the implementation of the primary adjustment method and considering results from alternative adjustment methods provides a sensible means to tackling sensitivity analyses in most settings. For studies using the IPCW method, key modifications might include changing the covariates included in the WD model or changing the functional form of covariates. Sensitivity analysis methods that explore the potential magnitude of bias due to unmeasured confounding should also be considered; see for example (47). For studies employing a RPSFTM, modifications might include changing the structural model assumed, or performing the analysis both with and without re-censoring applied (19). Importantly, allowing the AF to differ between treatment switchers and experimental arm patients (e.g. AF assumed to be 20% smaller in treatment switchers) is an intuitive approach for testing the sensitivity of results to the common treatment effect assumption.

One interesting discovery from the systematic review was the difference in the quality of reporting between published articles and NICE submissions. In particular, reporting on the RPSFTM tended to be more comprehensive in published articles than in NICE submissions, whereas the opposite trend was observed for the IPCW method. This pattern of reporting was also evident among the four trials represented in both a published article and a NICE TA. Given both the space constraints imposed on published articles and the fact that NICE submissions are prepared solely for the purpose of HTA decision-making, the improved reporting on the RPSFTM in published articles is somewhat counter-intuitive. Discounting chance differences due to the small number of included studies, one possible explanation for this observation is that several of the published articles involving the RPSFTM were secondary papers devoted entirely to the analysis of OS from the relevant trial.

In this review article we have focused primarily on reporting deficits rather than deficits in the adjustment methods themselves. Although we believe requirements for thorough reporting can only improve use of the methods, particularly since it encourages appropriate implementation and consideration of underlying assumptions, clearly thorough reporting on its own cannot ensure valid results. Choosing an appropriate class of adjustment method given the specifics of a trial is another critical step in producing appropriate conclusions, and to this end we would direct readers to the detailed framework provided by NICE on factors to consider when selecting a method of adjustment (48). In many trials, particularly those involving high switching proportions, ultimately a reliance on

unverifiable assumptions may make it difficult to draw definitive conclusions from a switching-adjusted analysis. Consequently, careful consideration should be given to the advantages and disadvantages of allowing for treatment switching during trial design.

During the review process we developed a checklist of reporting recommendations for switching-adjusted analyses in oncology trials. Designed to promote transparency and facilitate accurate assessments of validity, it is hoped these recommendations can be applied in practice to improve the quality and acceptance of future switching-adjusted analyses. In the context of HTA, adherence with the reporting recommendations will enable more informed decision-making on the cost-effectiveness of new treatments. Ultimately, this should lead to better decisions around the allocation of scarce health-care resources.

A limitation of this review is the small number of studies that satisfied inclusion criteria. Although this means that summary statistics describing the percentage of studies failing to report on some aspect of the analysis may be imprecise, we do not think this detracts from the overall message that reporting is currently inadequate. Another limitation of the review is that for feasibility data were only extracted from the 'Clinical Effectiveness' section of the initial submission to NICE. It is possible that additional details on the switching-adjusted analysis may have been provided in other sections of the submission or clarified during later correspondence with NICE. Another limitation is that we did not consider all possible methods for adjusting for treatment switching in this paper, for example iterative parameter estimation and structural nested models. Finally, our review focused on simple or "direct" treatment switching from the control to the experimental treatment, whereas in practice other types of switching is possible, for example "indirect" switching to a third treatment.

CONCLUSIONS

Despite the importance of switching-adjusted treatment effect estimates to HTA decision-making, the quality of reporting around the implementation of the RPSFTM and IPCW methods in published articles and industry submissions to NICE was generally poor. Based on these findings, it seems plausible that the acceptance of these methods of adjustment within the HTA context could be enhanced by adhering to the reporting recommendations presented in this article.

REFERENCES

1. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials--an economic evaluation context: methods, limitations, and recommendations. Medical Decision Making. 2014;34(3):387-402.
2. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013 [Available from: https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781.
3. Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. New York: Oxford University Press; 2006.
4. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada 2017 [Available from: https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf.
5. Watkins C, Huang X, Latimer N, Tang Y, Wright EJ. Adjusting overall survival for treatment switches: commonly used methods and practical application. Pharmaceutical Statistics. 2013;12(6):348-57.
6. White IR. Uses and limitations of randomization-based efficacy estimators. Statistical Methods in Medical Research. 2005;14(4):327-47.
7. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received: is it really an option? Statistics in Medicine. 1991;10(10):1595-605.
8. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics - Theory and Methods. 1991;20(8):2609-31.
9. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56(3):779-88.
10. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting for treatment switching in randomised controlled trials - A simulation study and a simplified two-stage method. Statistical Methods in Medical Research. 2017;26(2):724-51.
11. Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. BMC Medical Research Methodology. 2011;11:4.

12. Latimer NR, Abrams KR, Lambert PC, Morden JP, Crowther MJ. Assessing methods for dealing with treatment switching in clinical trials: A follow-up simulation study. Statistical Methods in Medical Research. 2018;27(3):765-84.
13. National Institute for Health and Care Excellence. Sunitinib for the treatment of gastrointestinal stromal tumours: TA 179: NICE; 2009.
14. National Institute for Health and Care Excellence. Pazopanib for the first-line treatment of advanced renal cell carcinoma: TA 215: NICE; 2010.
15. Latimer NR. Treatment switching in oncology trials and the acceptability of adjustment methods. Expert Review of Pharmacoeconomics and Outcomes Research. 2015;15(4):561-4.
16. Henshall C, Latimer NR, Sansom L, Ward RL. Treatment switching in cancer trials: issues and proposals. International Journal of Technology Assessment in Health Care. 2016;32(3):167-74.
17. White IR, Babiker AG, Walker S, Darbyshire JH. Randomization-based methods for correcting for treatment changes: examples from the Concorde trial. Statistics in Medicine. 1999;18(19):2617-34.
18. White IR, Walker S, Babiker A. strbee: Randomization-based efficacy estimator. Stata Journal. 2002;2(2):140-50.
19. Latimer NR, White IR, Abrams KR, Siebert U. Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times? Statistical Methods in Medical Research. 2018.
20. Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. Journal of the American Statistical Association. 2001;96(454):440-48.
21. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran M, Berry D, editors. Statistical models in epidemiology: the environment and clinical trials. New York: Springer; 1999. p. 95-134.
22. Dodd S, Williamson P, White IR. Adjustment for treatment changes in epilepsy trials: a comparison of causal methods for time-to-event outcomes. Statistical Methods in Medical Research. 2019;28(3):717-33.
23. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research. 2013;22(3):278-95.
24. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. Available from www.cochrane-handbook.org: The Cochrane Collaboration; 2011.

25. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Dore C, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. Journal of the American Medical Association. 2017;318(23):2337-43.
26. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. Statistics in Medicine. 1999;18(15):1903-42.
27. Colleoni M, Giobbie-Hurder A, Regan MM, Thurlimann B, Mouridsen H, Mauriac L, et al. Analyses adjusting for selective crossover show improved overall survival with adjuvant letrozole compared with tamoxifen in the BIG 1-98 study. Journal of Clinical Oncology. 2011;29(9):1117-24.
28. Demetri GD, Garrett CR, Schoffski P, Shah MH, Verweij J, Leyvraz S, et al. Complete longitudinal analyses of the randomized, placebo-controlled, phase III trial of sunitinib in patients with gastrointestinal stromal tumor following imatinib failure. Clinical Cancer Research. 2012;18(11):3170-9.
29. Faivre S, Niccoli P, Castellano D, Valle JW, Hammel P, Raoul JL, et al. Sunitinib in pancreatic neuroendocrine tumors: updated progression-free survival and final overall survival from a phase III randomized study. Annals of Oncology. 2017;28(2):339-43.
30. Jin H, Tu D, Zhao N, Shepherd LE, Goss PE. Longer-term outcomes of letrozole versus placebo after 5 years of tamoxifen in the NCIC CTG MA.17 trial: analyses adjusting for treatment crossover. Journal of Clinical Oncology. 2012;30(7):718-21.
31. Latimer NR, Abrams KR, Amonkar MM, Stapelkamp C, Swann RS. Adjusting for the Confounding Effects of Treatment Switching-The BREAK-3 Trial: Dabrafenib Versus Dacarbazine. Oncologist. 2015;20(7):798-805.
32. Latimer NR, Amonkar MM, Stapelkamp C, Sun P. Adjusting for confounding effects of treatment switching in a randomized phase II study of dabrafenib plus trametinib in BRAF V600+ metastatic melanoma. Melanoma Research. 2015;25(6):528-36.
33. Latimer NR, Bell H, Abrams KR, Amonkar MM, Casey M. Adjusting for treatment switching in the METRIC study shows further improved overall survival with trametinib compared with chemotherapy. Cancer Medicine. 2016;5(5):806-15.
34. Matulonis UA, Harter P, Gourley C, Friedlander M, Vergote I, Rustin G, et al. Olaparib maintenance therapy in patients with platinum-sensitive, relapsed serous ovarian cancer and a BRCA mutation: Overall survival adjusted for postprogression poly(adenosine diphosphate ribose) polymerase inhibitor therapy. Cancer. 2016;122(12):1844-52.

35. Metzger Filho O, Giobbie-Hurder A, Mallon E, Gusterson B, Viale G, Winer EP, et al. Relative Effectiveness of Letrozole Compared With Tamoxifen for Patients With Lobular Carcinoma in the BIG 1-98 Trial. Journal of Clinical Oncology. 2015;33(25):2772-9.
36. Motzer RJ, Escudier B, Oudard S, Hutson TE, Porta C, Bracarda S, et al. Phase 3 trial of everolimus for metastatic renal cell carcinoma : final results and analysis of prognostic factors. Cancer. 2010;116(18):4256-65.
37. Regan MM, Neven P, Giobbie-Hurder A, Goldhirsch A, Ejlertsen B, Mauriac L, et al. Assessment of letrozole and tamoxifen alone and in sequence for postmenopausal women with steroid hormone receptor-positive breast cancer: the BIG 1-98 randomised clinical trial at 8.1 years median follow-up. Lancet Oncology. 2011;12(12):1101-8.
38. Sternberg CN, Hawkins RE, Wagstaff J, Salman P, Mardiak J, Barrios CH, et al. A randomised, double-blind phase III study of pazopanib in patients with advanced and/or metastatic renal cell carcinoma: final overall survival results and safety update. European Journal of Cancer. 2013;49(6):1287-96.
39. Verstovsek S, Gotlib J, Mesa RA, Vannucchi AM, Kiladjian JJ, Cervantes F, et al. Long-term survival in patients treated with ruxolitinib for myelofibrosis: COMFORT-I and -II pooled analyses. Journal of Hematology and Oncology. 2017;10(1):156.
40. National Institute for Health and Care Excellence. Dabrafenib for treating unresectable or metastatic BRAF V600 mutation-positive melanoma: TA 321: NICE; 2014.
41. National Institute for Health and Care Excellence. Pembrolizumab for treating advanced melanoma after disease progression with ipilimumab: TA 357: NICE; 2015.
42. National Institute for Health and Care Excellence. Ruxolitinib for treating disease-related splenomegaly or symptoms in adults with myelofibrosis: TA 386: NICE; 2015.
43. National Institute for Health and Care Excellence. Crizotinib for untreated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer: TA 406: NICE; 2016.
44. National Institute for Health and Care Excellence. Everolimus for advanced renal cell carcinoma after previous treatment: TA 432: NICE; 2009.
45. National Institute for Health and Care Excellence. Regorafenib for previously treated unresectable or metastatic gastrointestinal stromal tumours: TA 488: NICE; 2017.
46. National Institute for Health and Care Excellence. Tivozanib for treating advanced renal cell carcinoma: TA 512: NICE; 2017.

47. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology (Cambridge, Mass). 2011;22(1):42-52.
48. Latimer NR, Abrams KR. NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching. Available from http://www.nicedsu.org.uk.; 2014.

Table 1. Recommendations for the reporting of switching-adjusted analyses

Item	Recommendation
All adjustment methods	
1	Provide results from an analysis unadjusted for treatment switching for comparison
2	Describe the treatment switching mechanism - who could switch and when
3	Detail the number of patients that switched, the number eligible to switch and when switching occurred
4	Give an overview of the data available for adjustment - what predictors and how frequently measured
5	State whether the chosen adjustment approach, including all model fitting steps, was pre-specified; if not, explain how the final model was selected*
6	Provide a statement around the plausibility of key assumptions (e.g. no unmeasured confounding for IPCW and common treatment effect for the RPSFTM)
7	Provide a visual comparison of observed and adjusted survival times
8	Report on sensitivity analyses showing the robustness of treatment effect estimates to violations of key assumptions
Inverse probability of censoring weights (IPCW)	
I.1	State whether unstabilized or stabilized weights were used
I.2	Detail the statistical procedure used to calculate weights (e.g. pooled logistic regression**, Cox model)
I.3	State the portion of data used in the WD model including time-varying predictors (e.g. post-progression data only)
I.4	Describe the extent of and the method used to address missing data on predictors in the WD model(s)
I.5	Present parameter estimates and associated measures of precision from the WD model(s)
I.6	Summarize the distribution of weights and state whether values were truncated
I.7	Detail the FO model, including estimation method (e.g. robust variance estimation) and baseline variables adjusted for
Rank preserving structural failure time model (RPSFTM)	
R.1	State and justify the structural model assumed (e.g. as treated, ever treated)
R.2	State the metric used for g-estimation (e.g. log-rank test), including baseline variables for adjustment where applicable
R.3	State the grid-search algorithm used
R.4	Plot g-estimation results to show that the estimation process has worked well
R.5	Present the estimated acceleration factor and its confidence interval
R.6	Compare counterfactual survival times between randomized groups in a Kaplan-Meier plot
R.7	Detail the FO model, including method for calculating a CI around the estimated treatment effect (e.g. retain ITT p-value, bootstrapping) and baseline variables adjusted for
R.8	Present results both with and without re-censoring applied

Abbreviations: AFT, accelerated failure time; CI, confidence interval; FO, final outcomes; IPCW, inverse probability of censoring weights; ITT, intention to treat; RPSFTM, rank preserving structural failure time model; WD, weight determining.

* Given the complexity of the methods, it may not always be feasible to fully pre-specify without consideration of the actual data collected or the performance of the models.

** Including the width of the discrete time interval and how time was adjusted for.

Table 2. Key characteristics of included trials

Characteristic	Published articles (n=13)	NICE TAs (n=8)
Number of participants: median (range)	416 (108, 5187)	326 (199, 517)
Percentage of treatment switchers in control arm: median (range)	64 (23, 87)	64 (40, 88)
Described the conditions under which switching was permitted	12 (92)	8 (100)
Provided treatment effect estimate* and CI from standard ITT analysis	12 (92)	8 (100)
Switching adjustment approach pre-specified		
Yes	0 (0)	2 (25)
No	8 (62)	1 (13)
Not stated	5 (38)	5 (63)
Provided justification for chosen adjustment method(s)	7 (54)	6 (75)
Provided a visual comparison of observed and adjusted survival times	6 (46)	5 (63)
Used the RPSFTM	9 (69)	6 (75)
Used IPCW	6 (46)	4 (50)
Used both the RPSFTM and IPCW	2 (15)	2 (25)

Values reported are numbers (percentages) unless otherwise indicated.
Abbreviations: CI, confidence interval; ITT, intention to treat; TA, technology appraisal.
* For example a hazard ratio or acceleration factor.

Table 3. Reporting on the use of the RPSFTM

Characteristic: n (%)	Published articles (n=9)	NICE TAs (n=6)
Described the structural model used	6 (67)	2 (33)
Stated metric used to demonstrate equivalence during g-estimation	5 (56)	2 (33)
Described the grid-search algorithm	0 (0)	1 (17)
Plotted g-estimation results	0 (0)	0 (0)
Presented estimated acceleration factor and 95% CI from g-estimation	2 (22)	1 (17)
Re-censoring applied		
Yes	5 (56)	2 (33)
Not stated	4 (44)	4 (67)
Compared counterfactual survival times between randomized groups	3 (33)	0 (0)
Method used for calculating confidence intervals in the final outcomes model		
Retain ITT p-value	5 (56)	1 (17)
Bootstrapping	1 (11)	1 (17)
Not stated	2 (22)	4 (67)
No CI reported	1 (11)	0 (0)
Stated the variables adjusted for in the final outcomes model	2 (22)	1 (17)
Assessed the sensitivity of results to the common treatment effect assumption	0 (0)	0 (0)

Abbreviations: CI, confidence interval; ITT, intention to treat; RPSFTM, rank preserving structural failure time model; TA, technology appraisal.

Table 4. Reporting on the use of the IPCW method

Characteristic: n (%)	Published articles (n=6)	NICE TAs (n=4)
Type of weight determining model		
Cox proportional hazards model	1 (17)	0 (0)
Pooled logistic regression	2 (33)	3 (75)
Not stated	3 (50)	1 (25)
Listed all covariates considered for the calculation of weights	1 (17)	3 (75)
Described the frequency of measurements for time-varying covariates	0 (0)	3 (75)
Reported extent of missing data on covariates and how missing data were handled	0 (0)	2 (50)
Used stabilized weights		
Yes	1 (17)	3 (75)
Not stated	5 (83)	1 (25)
Described the distribution of weights	1 (17)	2 (50)
Presented coefficient estimates and CIs from models used to calculate weights	0 (0)	2 (50)
Estimation method for calculating CI in the final outcomes model		
Robust variance estimation	2 (33)	0 (0)
Bootstrapping	1 (17)	0 (0)
Not stated	3 (50)	4 (100)
Stated the variables adjusted for in the final outcomes model	1 (17)	2 (50)

Abbreviations: CI, confidence interval; IPCW; inverse probability of censoring weights; TA, technology appraisal.

APPENDIX A: ELABORATION OF REPORTING RECOMMENDATIONS

All adjustment methods

Item 1: Provide results from an analysis unadjusted for treatment switching for comparison. This item is included to highlight the impact of adjustment on treatment effect estimates. Unadjusted results might be in the form of a hazard ratio, acceleration factor or log rank test p-value from an intention to treat (ITT) analysis. The statistical method and baseline variables adjusted for in producing the ITT treatment effect estimate should also be detailed.

Item 2: Describe the treatment switching mechanism - who could switch and when. A clear description of which patients were eligible to switch treatments and at what time-point during the trial switching was permitted gives an indication of which adjustment methods are potentially applicable (1).

Item 3: Detail the number of patients that switched, the number eligible to switch and when switching occurred. The extent and timing of treatment switching highlights its potential impact on ITT estimates, with higher switching proportions and shorter durations to switching typically associated with greater impact. These quantities are also useful in judging the appropriateness of alternative adjustment methods. For example, the IPCW method is known to be prone to bias when there are few non-switchers among patients eligible to switch (2).

Item 4: Give an overview of the data available for adjustment - what predictors and how frequently measured. Like items 1 and 2, this item gives an indication of which adjustment methods are potentially applicable. The no unmeasured confounding assumption of the IPCW method, for example, may be deemed inappropriate in trials involving few or sparsely collected predictors.

Item 5: State whether the chosen adjustment approach, including all model fitting steps, was pre-specified; if not, explain how the final model was selected. Pre-specification of statistical methods is recommended in randomised trials in order to minimise bias, maintain nominated type I error rates and avoid concerns over selective reporting (3). Given the complexity of switching adjustment methods, it may not always be feasible to follow pre-specified methods without consideration of the actual data collected. In this case, rationale should be provided for how the final model and its method of implementation (e.g. predictors included, structural model assumed) was chosen.

Item 6: Provide a statement around the plausibility of key assumptions (e.g. no unmeasured confounding for IPCW and two-stage approaches, and common treatment effect for the RPSFTM). A critical reporting item for all switching-adjusted analyses, the plausibility of key assumptions should be justified on both statistical and clinical grounds (1).

Item 7: Provide a visual comparison of observed and adjusted survival times. A Kaplan-Meier plot of observed and adjusted survival times in the control group is helpful for visualizing the impact of adjustment on estimated survival.

Item 8: Report on sensitivity analyses showing the robustness of treatment effect estimates to violations of key assumptions. As all switching adjustment methods rely on unverifiable assumptions, sensitivity analyses should be undertaken to assess the robustness of findings to plausible alternative assumptions. If treatment effect estimates can be shown to be similar across a range of sensitivity analyses, this will increase confidence in results. Sensitivity analyses may involve modifying the implementation of the primary adjustment method (e.g. relaxing the common treatment effect assumption of the RPSFTM or changing the predictors used in IPCW) or considering results from alternative adjustment methods. In the case of the IPCW method, the potential magnitude of bias due to unmeasured confounding should also be explored.

Inverse probability of censoring weights

Item I.1: State whether unstabilized or stabilized weights were used. Stabilized weights are generally recommended for the IPCW method (4), but the procedure may also be applied with unstabilized weights.

Item I.2: Detail the statistical procedure used to calculate weights (e.g. pooled logistic regression, Cox model). This item is intended to reduce uncertainty around the specification of the weight determining (WD) model(s) for the IPCW method. Where pooled logistic regression is employed, the width of the discrete time interval and how changes over time in the hazard of treatment switching were controlled for should be detailed; if a spline function of time was used, the type of spline and the procedure for placing knots should also be described.

Item I.3: State the portion of data used in the WD model including time-varying predictors (e.g. post-progression data only). To satisfy the requirement of positivity, the WD model including time-varying predictors should only be fitted to the portion of data where control group patients are at risk of switching. This reporting item can help to verify that the WD model was implemented appropriately.

Item I.4: Describe the extent of and the method used to address missing data on predictors in the WD model(s). Missing data on predictors in the WD model(s) could bias probability weights if inadequately addressed in the analysis, with the magnitude of bias likely to be greater with greater amounts of missing data. Hence both the extent of missing data and the method used to address it in the analysis (e.g. multiple imputation) should be stated. If there were no missing data on predictors, a statement to that effect should be provided.

Item I.5: Present parameter estimates and associated measures of precision from the WD model(s). Examining parameter estimates and associated measures of precision from the WD model(s) is a key step in demonstrating their appropriateness. Importantly, implausible parameter estimates or highly

inflated measures of precision (e.g. standard errors, confidence intervals) would cast doubt over the specification of the WD model(s) and hence the validity of ICPW treatment effect estimates.

Item I.6: Summarize the distribution of weights and state whether values were truncated. A summary of the distribution of weights, including the maximum obtained weight, is also important to consider in judging the appropriateness of the WD model(s), with extreme weights indicating potential problems with model specification. Where the weights are truncated at some upper limit to avoid extreme values, the method of truncation should also be described.

Item I.7: Detail the FO model, including estimation method (e.g. robust variance estimation) and baseline variables adjusted for. The final outcomes (FO) model should be fully described, including the statistical model used (Cox regression or pooled logistic regression), method of estimation (e.g. robust variance estimation) and the baseline variables adjusted for. This reporting item also offers a quick check that baseline variables used in the calculation of stabilized weights are controlled for in the FO model.

Rank preserving structural failure time model

Item R.1: State and justify the structural model assumed (e.g. as treated, ever treated). This is a critical reporting item since the structural model defines the underlying assumptions of the analysis.

Item R.2: State the metric used for g-estimation (e.g. log-rank test), including baseline variables for adjustment where applicable. This item entails fully describing the statistical test used to demonstrate equivalence between randomized groups in untreated survival times. Although it may be preferable to adopt the same model as used in the ITT analysis (5), any model could be chosen for the purpose of g-estimation.

Item R.3: State the grid search algorithm used. Different grid search algorithms are available for the estimating the acceleration factor during g-estimation, for example searching in fixed steps or using interval bisection. This reporting item, like items R.1 and R.2 above, is necessary for fully describing how the RPSFTM was implemented.

Item R.4: Plot g-estimation results to show that the estimation process has worked well. The performance of g-estimation should be assessed by plotting potential values for the acceleration factor against the chosen test statistic (6). If successful, the g-estimation procedure should identify a unique solution where the test statistic equals zero.

Item R.5: Present the estimated acceleration factor and its confidence interval. For the as-treated and ever-treated structural models, the estimated acceleration factor from g-estimation conveys the effect of the experimental treatment on extending overall survival. The magnitude of this causal effect, along with its confidence intervals, should be scrutinised closely when judging the validity of the RPSTM.

Item R.6: Compare counterfactual survival times between randomized groups in a Kaplan-Meier plot. Assuming g-estimation is successful, counterfactual survival times should appear equivalently distributed across randomized groups. Differences in the distribution could indicate problems with g-estimation or the assumptions of the chosen structural model.

Item R.7: Detail the FO model, including method for calculating a CI around the estimated treatment effect (e.g. retain intention to treat p-value, bootstrapping) and baseline variables adjusted for. As well as explaining the process for generating counterfactual survival times, it is critical to detail how these times were analyzed in the FO model.

Item R.8: Present results both with and without re-censoring applied. In general, it will be informative to consider results from the RSPFTM both with and without re-censoring applied (7).

APPENDIX B: RECOMMENDATIONS FOR TWO STAGE ADJUSTMENT

Like the RPSFTM, the two-stage adjustment method (2) uses a potential outcomes framework to estimate counterfactual survival times that would have been observed had treatment switching not occurred. The first step in applying the method is to estimate the effect of treatment switching on extending overall survival (OS) in the control arm beyond some "secondary baseline", defined as a time-point where patients are at a similar stage of disease and in which switching cannot occur prior. Since treatment switching in oncology trials is usually only permitted following disease progression, the time of progression is a standard choice for the secondary baseline. Once defined, the effect of treatment switching beyond the secondary baseline can be estimated using an accelerated failure time (AFT) model, with switching treated as a time-dependent variable and with adjustment for predictors of switching and subsequent survival measured at the time of the secondary baseline. Once the acceleration factor (AF) due to switching has been estimated, the two-stage method proceeds in a similar manner to the RPSFTM. Revised survival times beyond the secondary baseline can be calculated using the same form of structural model as for the RPSFTM, but with U_i, T_{Ci} and T_{Ei} representing survival times beyond the secondary baseline rather than total survival times. Counterfactual survival times are then obtained by adding the time prior to the secondary baseline to these revised survival times. As with the RPSFTM, re-censoring may be applied at this stage to avoid informative censoring. The counterfactual survival times can then be included in a final outcomes (FO) model, with bootstrapping applied to the entire estimation procedure to obtain valid confidence intervals for the switching-adjusted treatment effect estimate (2).

As well as the existence of a suitable secondary baseline, the validity of the two-stage method relies on satisfying the no unmeasured confounding (NUC) assumption (2). To meet this assumption, all characteristics defined at the time of the secondary baseline predictive of both treatment switching and subsequent survival should be adjusted for in the AFT model. As the two-stage method was not designed to incorporate predictors measured after the secondary baseline, a further requirement for satisfying NUC is that there is no time-dependent confounding between the secondary baseline and the time of

treatment switching. This may not be realistic if switching frequently occurs much later than the secondary baseline. As with the IPCW method, clinical justification should be provided regarding the plausibility of NUC, with consideration given to both the predictors included in the AFT model and how soon treatment switching occurs after the secondary baseline. To produce valid treatment effect estimates, the AFT model for OS beyond the secondary baseline must also be correctly specified. This entails choosing an appropriate parametric distribution for survival, correctly specifying the functional form of predictors in the model, and, if necessary, addressing missing data on predictors under a plausible assumption about the reason for missing data. Examining coefficient estimates from the AFT model can be helpful in checking the suitability of its specification.

In addition to estimating the switching-adjusted effect of treatment, the two-stage method can be used to explore the appropriateness of other adjustment methods. For example, under the CTE assumption of the RPSFTM, one would expect the two-stage method to produce a similar treatment effect estimate for switching patients to the effect estimated for patients initially randomized to the experimental group. If these estimates differ substantially, the RPSFTM may be deemed inappropriate.

Based on the assumptions and model fitting procedure of two-stage adjustment, in Table B1 below we offer some suggestions on what should be reported following an analysis with this method.

Table B1. Recommendations for reporting on two-stage adjustment

Item	Recommendation
T.1	Summarize the distribution of time between the secondary baseline and treatment switching
T.2	State and justify the distribution assumed in the AFT model
T.3	Describe the extent of and the method used to address missing data on predictors in the AFT model
T.4	Present parameter estimates and associated measures of precision from the AFT model
T.5	Detail the FO model, including estimation method (e.g. bootstrapping) and baseline variables adjusted for
T.6	Present results both with and without re-censoring applied

Abbreviations: AFT, accelerated failure time; FO, final outcomes.

Elaboration of reporting recommendations for two-stage adjustment

Item T.1: Summarize the distribution of time between the secondary baseline and treatment switching. A key assumption of the two-stage method is that there is no time-dependent confounding between the secondary baseline and the time of treatment switching. Such an assumption may be unrealistic if switching occurs much later than the secondary baseline.

Item T.2: State and justify the distribution assumed in the AFT model. This item involves reporting the parametric distribution (e.g. Weibull, log-logistic) assumed in the AFT model for OS beyond the secondary baseline.

Item T.3: Describe the extent of and the method used to address missing data on predictors in the AFT model. Since missing data on predictors in the AFT model has the potential to bias estimation, both the extent of and the method used to address missing data should be detailed. If there were no missing data on predictors, a statement to that effect should be provided.

Item T.4: Present parameter estimates and associated measures of precision from the AFT model. Parameter estimates should be reported to demonstrate the suitability of the chosen AFT model. Particular attention should be given to the point estimate and associated measure of precision for treatment switching, as this describes the causal effect of the experimental treatment on extending OS and is the key quantity used in calculating counterfactual survival times.

Item T.5: Detail the FO model, including estimation method (e.g. bootstrapping) and baseline variables adjusted for. Once again the FO model should be fully detailed, including the statistical model used, method of estimation (e.g. bootstrapping entire estimation procedure) and baseline variables adjusted for.

Item T.6: Present results both with and without re-censoring applied. Consistent with the corresponding reporting item for the RPSFTM, it is generally informative to consider results both with and without re-censoring applied (7).

APPENDIX C: APPLICATION OF REPORTING RECOMMENDATIONS TO A CASE STUDY

The case study dataset for illustrating the reporting recommendations was generated using data simulation. As the aim of simulation was to produce a single realistic dataset for analysis and not to evaluate statistical properties over repeated samples, we provide only a very brief overview of the design here. Initially, baseline values for age, gender, a tumour biomarker, prognosis score, Eastern Cooperative Group Oncology (ECOG) score, and health related quality of life (HRQOL) were simulated for 300 patients. The patients were then allocated to an experimental or control group in the ratio 2:1 via simple randomisation. Using the same approach as in (8), a joint longitudinal model was then used to simulate OS times and time-dependent biomarker values every 21 days, with the hazard for death depending on the biomarker value at the corresponding time point, randomised group and baseline prognosis score. This model produces a treatment effect for OS that initially increases during the period of greatest hazard and then decreases with longer follow-up, as might be expected in real settings. Time to disease progression was then generated as a proportion of OS, with survival times administratively censored at 18 months. Next, treatment switching was introduced so that approximately 40% of control group patients switched within 42 days of disease progression. Importantly, the odds of switching increased with greater progression free survival and higher biomarker values at disease progression and decreased over the two 21-day periods following progression. Among switching patients, OS times were extended by applying the average treatment effect received by the experimental group reduced by 20%. Additional variables generated for the case study included time-dependent HRQOL and ECOG score at disease progression, the latter of which was strongly associated with (but not an independent predictor of) treatment switching. A summary of the variables in the case study dataset and their relationship with treatment switching is provided in Table C1.

Table C1. Variables in case study dataset and their association with treatment switching

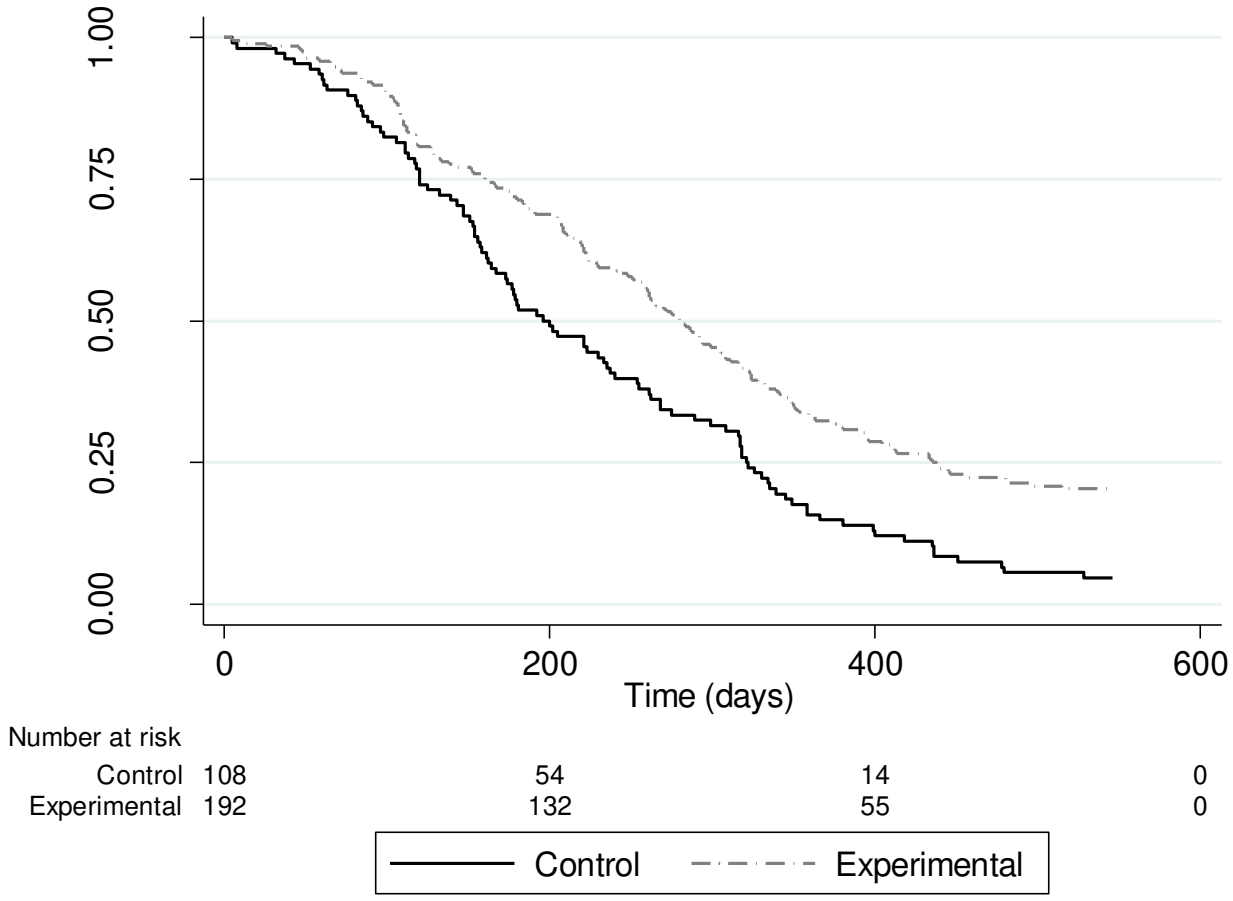
Variable	Categories	Frequency of time-varying measurements	Relation to probability of treatment switching
Baseline only			
Treatment group	Experimental Control	-	-
Gender	Female Male	-	Unrelated
Prognosis score	Good Poor	-	Related (through effect on time to disease progression)
Age	-	-	Unrelated
Disease progression only			
Time to disease progression	-	-	Predictor
Baseline and time-varying			
ECOG score	1 2 3 4	Disease progression only	Related (ECOG at baseline related through association with prognosis score; ECOG at disease progression strongly related)
Biomarker	-	Every 21 days	Predictor (value at disease progression)
HRQOL	-	Every 21 days	Unrelated
Time-varying only			
Time since disease progression	-	-	Predictor

Abbreviations: ECOG, Eastern Cooperative Oncology Group; HRQOL, health-related quality of life

All statistical analyses were performed using Stata 14.0 (StataCorp). In the chosen simulated dataset, 192 and 108 patients were randomized to the experimental and control arms, respectively. Baseline characteristics were generally well balanced between groups, although by chance the experimental arm had a higher percentage of patients with a poor prognosis (52.6% vs. 47.2%). As evident in Figure C1, OS was noticeably improved with the experimental treatment. According to an unadjusted Cox model, the experimental treatment reduced the hazard of death relative to control by 41% (ITT hazard ratio (HR) = 0.59; 95% confidence interval (CI) 0.46-0.76). With adjustment for baseline prognosis score,

ECOG score, biomarker value and HRQOL, the treatment effect became stronger than in the unadjusted analysis (adjusted ITT HR = 0.50; 95% CI 0.38-0.65) **[Reporting recommendations: item 1]**, mostly due to the observed imbalance in the baseline prognosis score. Of course, due to treatment switching, both these estimates may understate the true benefit of the experimental treatment.

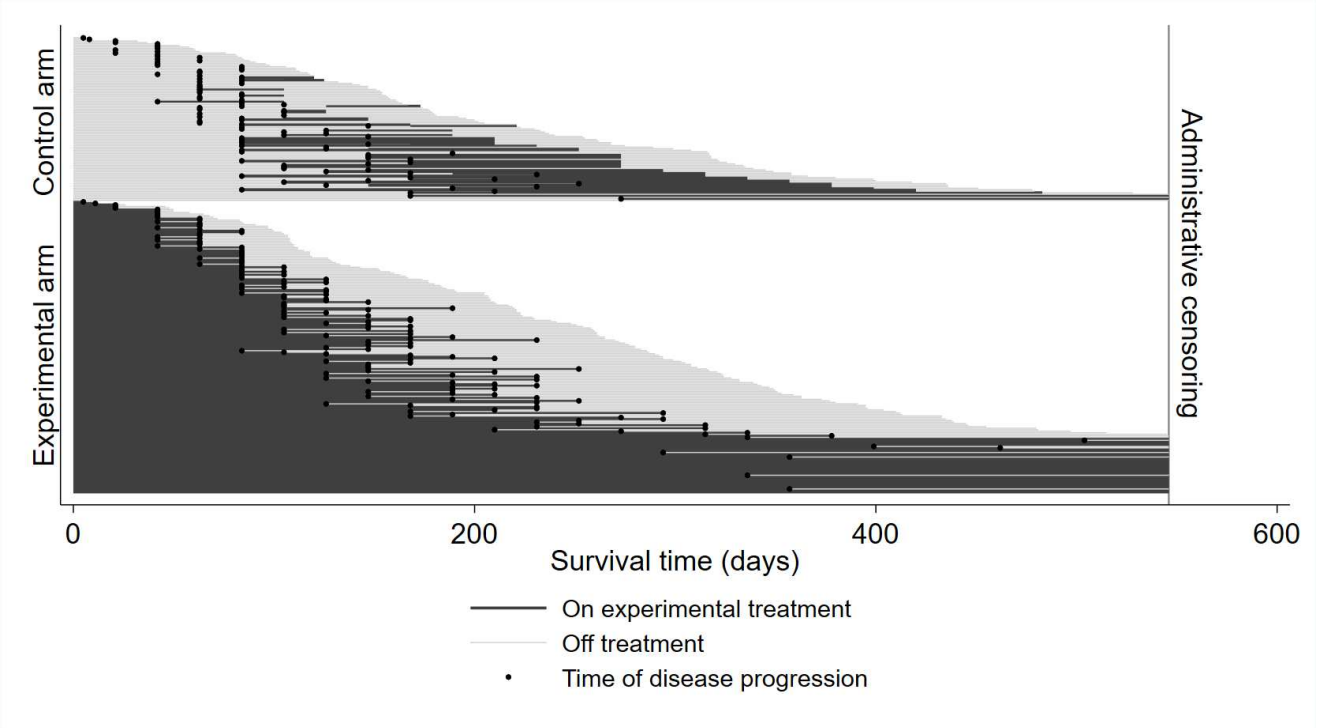
Figure C1. Kaplan Meier plot for overall survival



As per the simulation design, control group patients could switch to the experimental treatment within 42 days of disease progression **[item 2]**. As illustrated in Figure C2, 49/108 control group patients (45.4%) switched treatments, including 39 that switched immediately upon disease progression and 10 that switched either 21 or 42 days later (median time from randomisation to switching of 126 days). Only 5 control group patients were ineligible to switch treatments; two died in the first 21 days of the trial and three had not progressed by the time of administrative censoring **[item 3]**. An overview of the

data available for performing switching-adjusted analyses is provided in Table C1 **[item 4]**. For illustration purposes, we assume that the implementation of each switching-adjustment method, as described below, was entirely pre-specified **[item 5]**.

Figure C2. Time spent on the experimental treatment by randomized group



Inverse probability of censoring weights

Stabilized weights **[item I.1]** were estimated using pooled logistic regression based on 21-day interval data, with a restricted cubic spline containing three internal knots used to control for changes in the hazard of switching over time; these knots were spaced equally according to percentiles of observed switching times **[item I.2]**. For the WD model involving time-varying predictors, only data within 42 days of disease progression, where patients were at risk of switching, were incorporated **[item I.3]**. The odds of not switching was modelled according to prognosis score, biomarker value, ECOG score (treated as continuous to avoid convergence issues), HRQOL and time since disease progression. We imagine these variables, all fully observed **[item I.4]**, were selected based on expert clinical input and that the NUC assumption holds given these variables. Parameter estimates from the WD models are

presented in Table C2 **[item I.5]**. In the model including time-varying predictors, ECOG scores at disease progression were found to be highly related to not switching treatments (odds ratio = 24.8); this is unsurprising given that 46/49 switching patients had an ECOG score of 1 at disease progression, compared to just 3/54 for non-switchers. In contrast, other predictors appeared to have little impact on the odds of remaining on the control treatment. Stabilized (untruncated) weights ranged between 0.1 and 11.1, with a coefficient of variation of 0.54 **[item I.6]**. Taken together with the extremely wide CIs observed for several of the predictors (see Table C2), it seems the IPCW approach was rather unstable, with results sensitive to a handful of patients with large weights (specifically the three patients with an ECOG value of 1 at disease progression who remained on the control treatment). On statistical grounds this raises concerns about the plausibility of the method **[item 6]**.

Table C2. Results from weight determining models

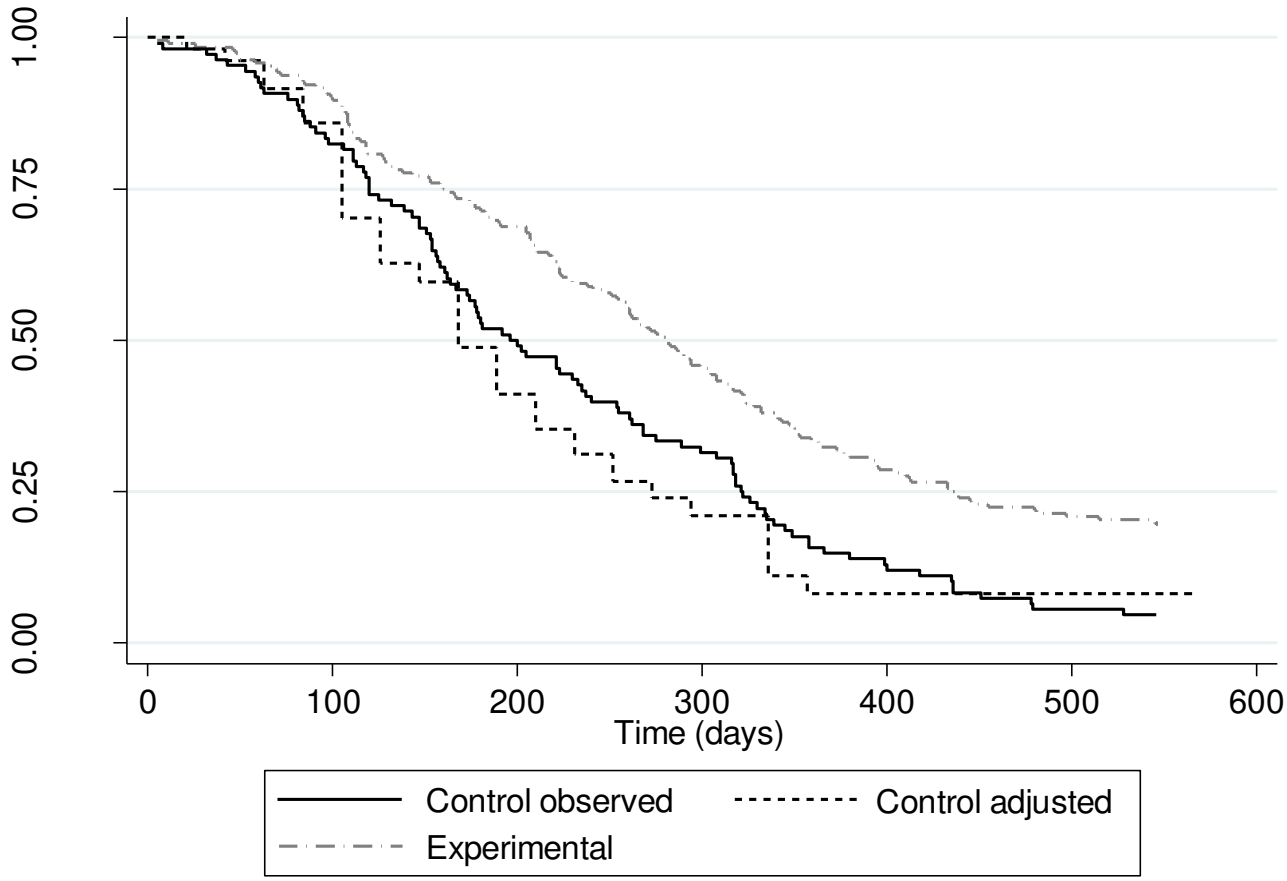
Characteristic	Odds ratio (95% CI)	
	Baseline predictors only (for numerator of stabilized weights)	Baseline and time-varying predictors (for denominator of stabilized weights)
Baseline prognosis score (poor vs. good)	14.91 (4.61 to 48.2)	3.00 (0.00 to 2306)
Baseline biomarker value	0.89 (0.63 to 1.24)	1.20 (0.51 to 2.84)
Baseline ECOG score	0.97 (0.51 to 1.84)	0.66 (0.19 to 2.25)
Baseline HRQOL	9.76 (0.43 to 222)	0.08 (0.00 to 30.0)
Time since progression (at this visit vs. 1 or 2 visits ago)		0.88 (0.22 to 3.60)
Current biomarker value		1.10 (0.59 to 2.02)
ECOG score at disease progression		24.76 (4.93 to 124)
Current HRQOL		0.53 (0.00 to 294)

Abbreviations: ECOG, Eastern Cooperative Oncology Group; HRQOL, health-related quality of life

Pooled logistic regression based on 21-day interval data was also used for the weighted FO model, with a restricted cubic spline with three internal knots (spaced equally according to percentiles of observed death times) applied to control for changes in the hazard of death over time. In the model, robust

variance estimation was used to account for the uncertainty in the stabilized weights, while adjustment was made for prognosis score, baseline biomarker value, baseline ECOG score (treated as continuous) and baseline HRQOL **[item I.7]**. The FO model produced a switching-adjusted HR of 0.39 (95% CI 0.23-0.64), substantially lower than the ITT estimate. For a visual comparison of observed and adjusted survival times from the IPCW method, see Figure C3 below **[item 7]**.

Figure C3. Observed and adjusted survival times from the IPCW base case analysis using unstabilized weights*



* Survival curve plotted using unstabilized weights, as any analysis involving stabilized weights needs to control for the baseline variables included in the weighting models.

Rank preserving structural failure time model

For the RPSFTM, counterfactual survival times were calculated assuming an ever-treated structural model **[item R.1]**, with a long-rank test **[item R.2]** using interval bisection **[item R.3]** employed for g-

estimation. These calculations were performed using the Stata module *strbee* (9). As shown in Figure C4, g-estimation produced a single unique solution for $\psi = -\log(\text{AF})$ where the Z-statistic for the log-rank test equalled zero [$\psi = -0.54$, **item R.4**]. The corresponding AF for this estimate was 1.72 (95% CI 1.38-2.12) **[item R.5]**, indicating that the experimental treatment extended OS by 72% compared to control. Figure C5 displays counterfactual survival times assuming both randomized groups received only the control treatment **[item R.6]**. When the CTE and perfect randomization assumptions of the RPSFTM are met, one would expect equal counterfactual survival times across randomized groups, which was not the case here. Instead, counterfactual survival times were somewhat divergent, with poorer survival in the experimental group in the first 100 days. This casts some doubt over the plausibility of underlying assumptions **[item 6]**.

Figure C4. Plot of g-estimation results

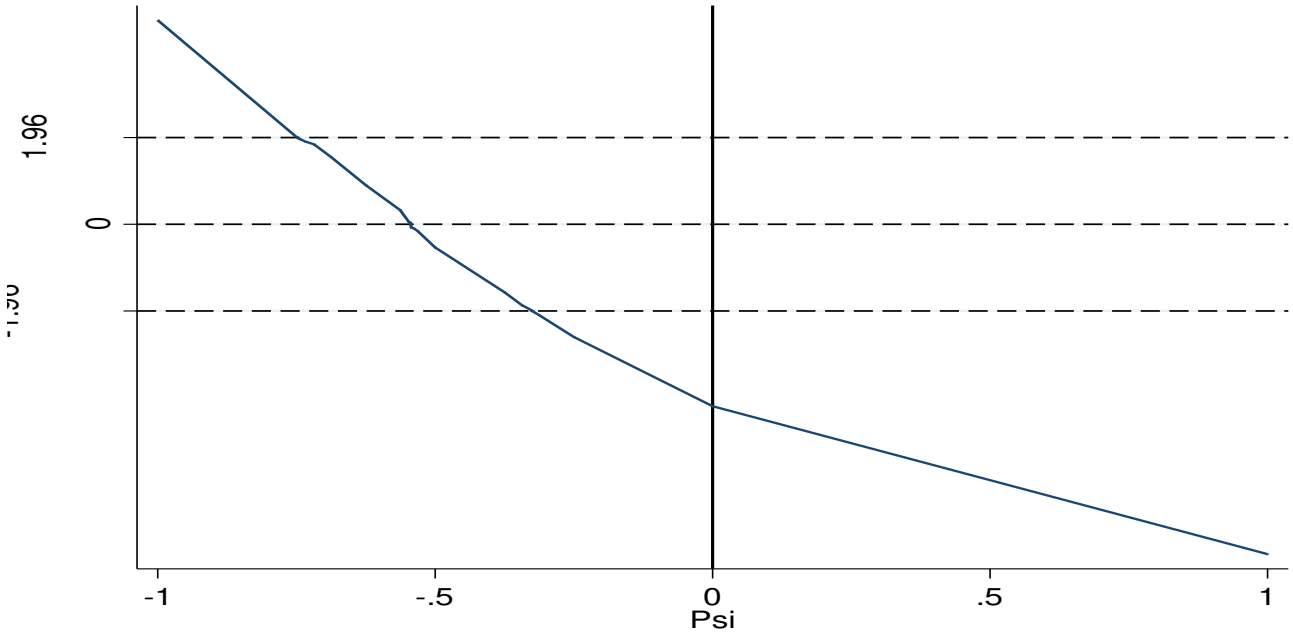
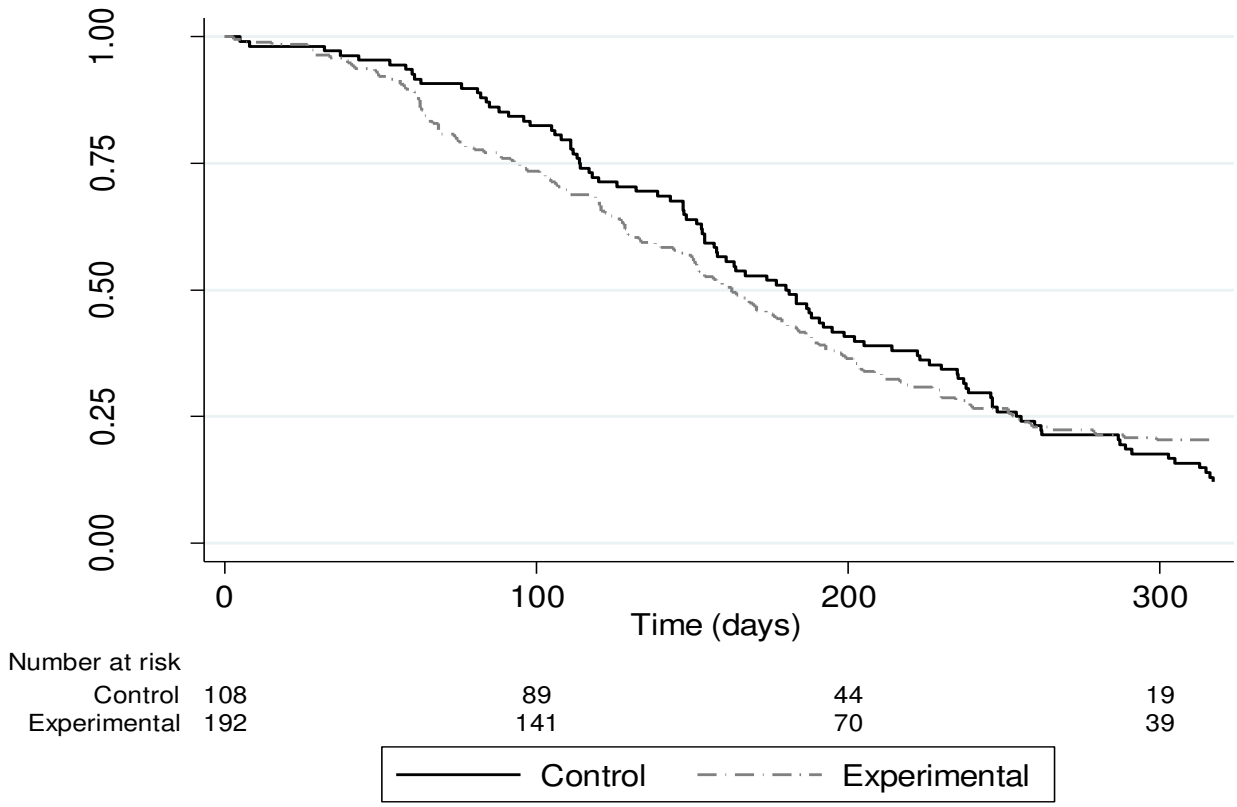
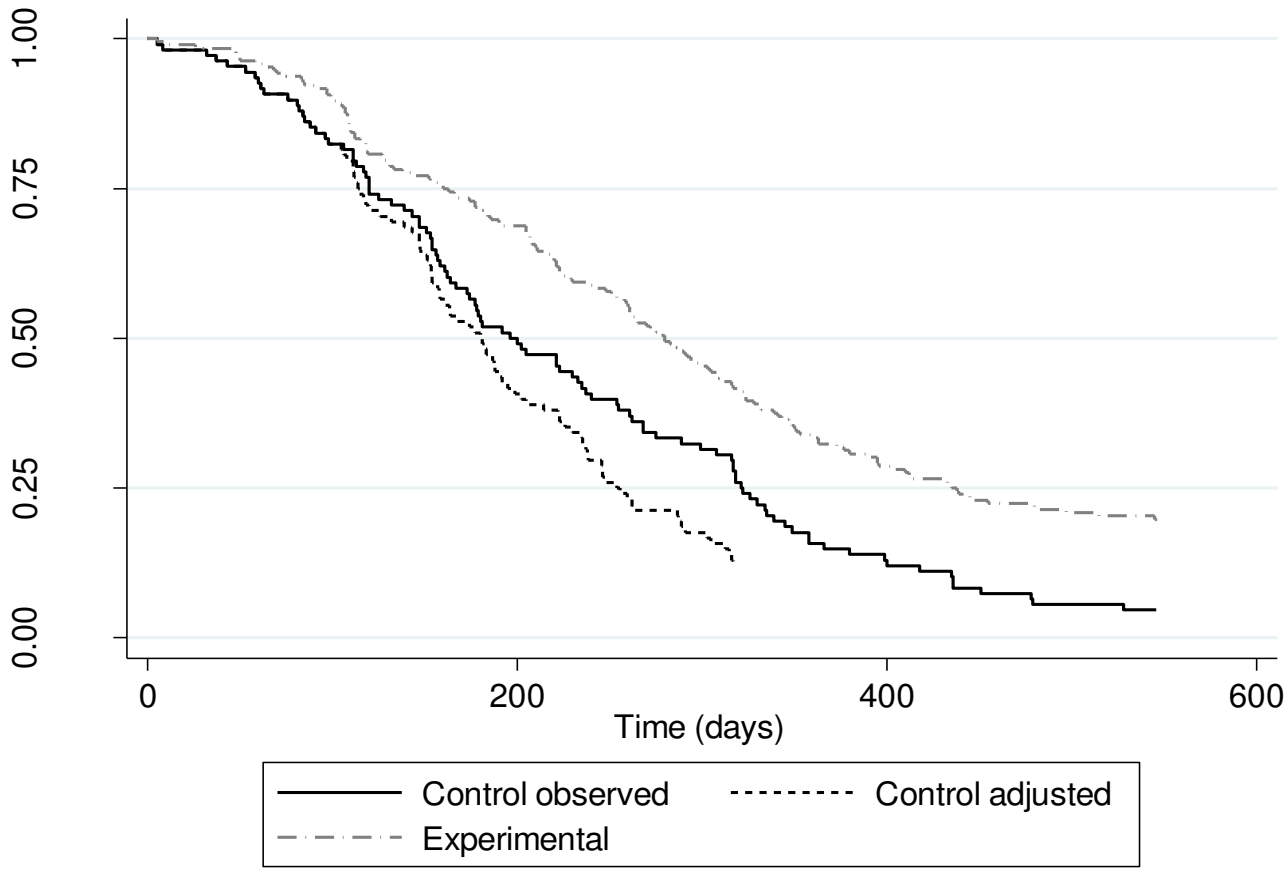


Figure C5. Comparison of counterfactual survival times between groups (assuming only control treatment given)



The switching-adjusted HR for OS in the FO model, an unadjusted Cox model, was estimated to be 0.44 with re-censoring applied (95% CI 0.30-0.65; calculated by retaining the ITT p-value) and 0.46 without re-censoring (95% CI 0.31-0.66) **[items R.7 & R.8]**. A comparison of observed and re-censored adjusted survival times from the RPSFTM is presented in Figure C6 **[item 7]**. As expected, the survival experience of the control arm was poorer after switching had been adjusted for, while re-censoring was associated with a substantial loss of longer-term survival information.

Figure C6. Observed and adjusted survival times from the RPSFTM base case analysis



Two-stage adjustment

Of the 49 control group patients that switched treatments, 39 switched immediately upon disease progression, while 8 and 2 patients switched 21- and 42-days following progression, respectively **[item T.1]**. In fitting the two-stage model, the AF due to switching was estimated from a Weibull AFT model **[item T.2]**, with adjustment for prognosis score and the following fully-observed predictors **[item T.3]** at disease progression: biomarker value, ECOG score (treated as continuous), HRQOL and time to disease progression. As with the IPCW method, we imagine these variables were informed by expert clinical input and that the NUC assumption holds given these variables. Parameter estimates from the AFT model are presented in Table C3 **[item T.4]**. As shown in the table, the Weibull model estimated an AF due to treatment switching of 1.86 (95% CI 1.15-3.01), with time to disease progression the only other statistically significant predictor of post-progression survival in the model. It should be noted that

the AF of 1.86 was similar to the corresponding estimate from the RPSFTM (1.72), lending some credibility to the CTE assumption. Overall the Weibull AFT model produced sensible parameter estimates, which, in combination with the choice of predictors and the short duration of time between disease progression and switching, suggests that the assumptions of two-stage adjustment were reasonable for these data **[item 6]**.

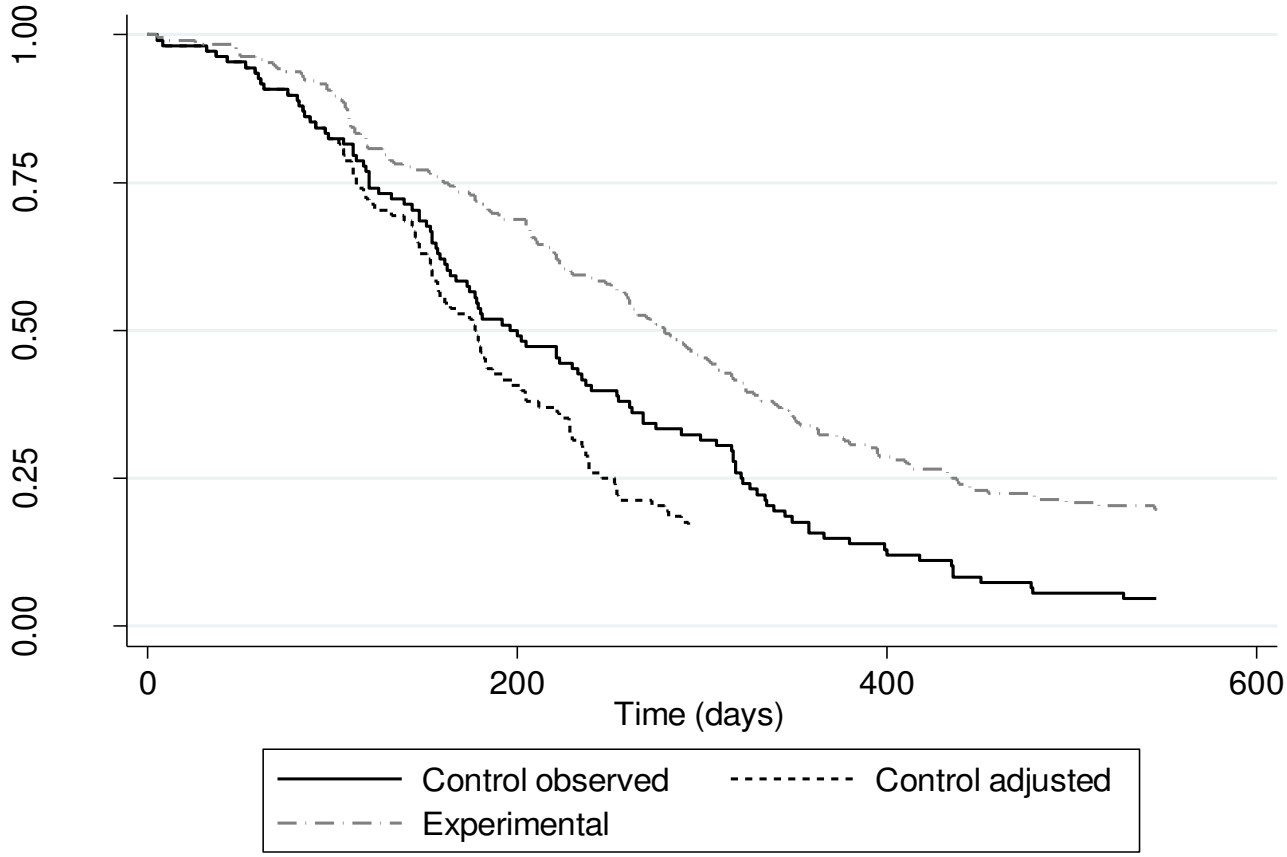
Table C3. Results from accelerated failure time model

Characteristic	Acceleration factor (95% CI)
Treatment switching	1.86 (1.15 to 3.01)
Baseline prognosis score (poor vs. good)	0.83 (0.42 to 1.63)
Biomarker value at disease progression	1.02 (0.96 to 1.08)
ECOG score at disease progression	1.02 (0.76 to 1.38)
HRQOL at disease progression	0.47 (0.16 to 1.38)
Time to disease progression (per 100 days) *	1.05 (1.01 to 1.09)

Abbreviations: ECOG, Eastern Cooperative Oncology Group; HRQOL, health-related quality of life
* Parameter estimate describes the effect of a 10-day increase in time to disease progression

Following the re-censoring of counterfactual survival times in the control group, an unadjusted Cox model was fitted to the data, producing a switching-adjusted HR for OS of 0.44 (95% CI 0.30-0.62; calculated by bootstrapping the entire estimation procedure 1000 times and applying the percentile method **[item T.5]**). The switching-adjusted HR was also 0.44 without re-censoring applied (95% CI 0.31-0.62) **[item T.6]**. A visual comparison of observed and re-censored adjusted survival times is provided in Figure C7 **[item 7]**.

Figure C7. Observed and adjusted survival times from the two-stage adjustment base case analysis



Sensitivity analyses

A variety of sensitivity analyses were undertaken to investigate the robustness of treatment effect estimates to the assumptions of the adjustment methods **[item 8]**. For IPCW, changing the specification of the WD and FO models produced switching-adjusted HRs that ranged between 0.39 and 0.47 (Table C4). Increasing the number of knots in spline functions or adding in variables unrelated to switching and OS had little impact on estimates, whereas the switching-adjusted HR noticeably increased when linear terms for time were used in the analysis. As (unweighted) pooled logistic regression more accurately reproduced the ITT HR from Cox regression when spline rather than linear terms for time were used, our preference here is for the models including spline terms. It is also interesting to note the worsening in model fit, as indicated by the Akaike information criterion, when ECOG scores were excluded from the WD and FO models. Despite this, the exclusion greatly reduced the maximum

stabilized weight and improved the precision of the switching-adjusted HR. With knowledge that the data generation procedure for treatment switching was not dependent on ECOG scores (i.e. NUC holds without ECOG), this may be the preferred implementation of the IPCW method in Table C4.

Table C4. Sensitivity of treatment effect estimates for the IPCW method

Approach	Adjusted HR (95% CI)	Maximum weight [CV]	AIC*
(1) Base-case	0.39 (0.23 to 0.64)	11.1 [0.54]	105.8
(2) Same as 1 but using unstabilized weights	0.44 (0.27 to 0.71)	14.0 [0.52]	105.8
(3) Same as 1 but using 4 internal knots for spline functions of time	0.41 (0.26 to 0.64)	8.0 [0.44]	105.1
(4) Same as 1 but using linear terms (rather than spline functions) for time	0.47 (0.33 to 0.67)	4.8 [0.43]	106.9
(5) Same as 1 but including age and gender (both unrelated to switching & OS)	0.38 (0.23 to 0.62)	10.8 [0.52]	109.8
(6) Same as 1 but excluding ECOG scores at baseline and disease progression	0.44 (0.31 to 0.64)	3.0 [0.38]	125.8

Abbreviations: AIC, Akaike information criterion; CI, confidence interval; CV, coefficient of variation; ECOG, Eastern Cooperative Oncology Group; HR, hazard ratio; OS, overall survival
* AIC of weight determining model including both baseline and time-dependent covariates

The results of sensitivity analyses for the RPSFTM and two-stage approach are presented in Table C5. As shown in the table, switching-adjusted HRs for the RPSFTM ranged between 0.41 and 0.52 according to the exact method of implementation. The highest HR was obtained when switching patients were assumed to receive only 50% of the benefit of the experimental treatment, while the lowest HR was produced when an adjusted Cox model was used for g-estimation (mostly due to adjustment for prognostic variables in the FO model). As observed in the base-case RPSFTM, counterfactual survival times remained unequally distributed across randomized groups throughout the sensitivity analyses, including models relaxing the CTE assumption and with correction for baseline imbalances in the prognosis score (results not shown). This casts further doubt over the validity of the RPSFTM. For the two-stage approach, switching-adjusted HRs ranged from 0.41 to 0.45 across the sensitivity analyses, with the choice of parametric survival distribution having little impact on results. Consistent with findings for the IPCW method, the switching-adjusted HR following two-stage adjustment became more precise when ECOG scores were excluded from the estimation procedure.

Table C5. Sensitivity of treatment effect estimates for the RPSFTM and two-stage approaches

Approach	AF (95% CI)	Adjusted HR (95% CI)
Rank preserving structural failure time model		
(R1) Base-case with re-censoring applied	1.72 (1.38 to 2.12)	0.44 (0.30 to 0.65)
(R2) Base-case without re-censoring applied	1.72 (1.38 to 2.12)	0.46 (0.31 to 0.66)
(R3) Same as R1 but using an as-treated structural model	2.96 (1.72 to 4.63)	0.45 (0.31 to 0.66)
(R4) Same as R1 but using a Cox model for g-estimation*	1.73 (1.45 to 2.10)	0.41 (0.29 to 0.58)
(R5) Same as R1 but using a different grid search method**	1.72 (1.38 to 2.11)	0.45 (0.31 to 0.66)
(R6) Same as R1 but assuming treatment effect in switchers is 25% smaller	1.65 (1.35 to 2.00)	0.48 (0.34 to 0.68)
(R7) Same as R1 but assuming treatment effect in switchers is 50% smaller	1.56 (1.29 to 1.89)	0.52 (0.38 to 0.71)
Two-stage adjustment		
(T1) Base case with re-censoring applied	1.86 (1.15 to 3.01)	0.44 (0.30 to 0.62)
(T2) Base-case without re-censoring applied	1.86 (1.15 to 3.01)	0.44 (0.31 to 0.62)
(T3) Same as T1 but using log-logistic model	2.09 (1.30 to 3.37)	0.41 (0.29 to 0.62)
(T4) Same as T1 but using generalised gamma model	1.86 (1.15 to 3.01)	0.44 (0.30 to 0.62)
(T5) Same as T1 but including age and gender (both unrelated to switching)	1.85 (1.15 to 3.00)	0.45 (0.30 to 0.63)
(T6) Same as T1 but excluding ECOG score at disease progression	1.83 (1.22 to 2.73)	0.45 (0.32 to 0.60)

Abbreviations: AF, acceleration factor; CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; HR, hazard ratio
* Adjusted for baseline prognosis score, Eastern Cooperative Oncology Group score, health-related quality of life and biomarker value
** Grid search for psi from -2 to 0 in steps of 0.01

Discussion of results

As well as demonstrating the application of the reporting recommendations, the analyses of the case study dataset show how the recommendations can improve confidence in switching-adjusted estimates. In following the recommendations, we identified performance issues with the IPCW and RPSFTM approaches, including questionable fit of the WD model and non-equivalence of counterfactual survival times. Conversely, the two-stage method was supported by the short time duration between disease progression and treatment switching, and seemed to work well based on parameter estimates and the consistency of sensitivity analyses. That two-stage adjustment produced an equivalent switching-

adjusted HR to the base-case RPSFTM with re-censoring applied and IPCW excluding ECOG scores (our preferred specification for IPCW) provides further assurance on its suitability. Another interesting finding from the case study analysis was the sensitivity of results to the specification of each switching-adjustment approach. For example, switching-adjusted HRs varied between 0.39 and 0.47 for IPCW based on relatively small changes to its specification; such changes could have large impacts on final cost-effectiveness estimates. This reinforces the need to fully describe and justify the chosen modelling approach.

Importantly, since the case study dataset was generated using simulation, we did not explore the plausibility of model assumptions according to expert clinical opinion, as would be expected in practice (1). In justifying the NUC assumption, one should closely examine how treatment switching decisions were made in the trial and what information was available to guide these decisions. Of note, a characteristic can only confound if available to those responsible for making the switching decisions. Additionally, sensitivity analysis methods that explore the potential magnitude of bias due to unmeasured confounding should be considered; see for example (10). For the CTE assumption, the mechanism by which the experimental treatment works and its likely effectiveness at different stages of disease are key considerations.

Statistical code for analyses

The variables included in the base case analyses of the case study dataset are detailed in Table C6. Of note, the data were arranged in long format, with each row of the dataset corresponding to the start of a 21-day period for a given patient. A small snapshot of the data is provided in Table C7 to illustrate this setup.

Table C6. Variables in case study dataset

Variable	Variable name	Categorical values
Study ID	id	-
Time (in 21-day intervals)	time	-
Treatment group	trtrand	0 = control, 1 = experimental treatment
Time to death - continuous	deathtime	-
Death in 21-day interval	deathtdo	0 = no, 1 = yes
Death indicator	deathind	0 = alive, 1 = death
Time to disease progression	progtime	-
Disease progression indicator	progind	0 = no progression, 1 = progressed
Disease progression in previous 42 days	progtdc	0 = no, 1 = yes
Time to treatment switching	xotime	-
Switched to experimental treatment indicator	xoind	0 = no switch, 1 = switched
Switched to experimental treatment in 21-day interval	xotdoipcw	0 = no, 1 = yes
Time to discontinuing experimental treatment	disconexp	-
Prognosis score	PROGNOSISb	
ECOG score*	ECOGb / ECOGtdc	
Biomarker*	CEAb / CEAtdc	
HRQOL*	HRQLb / HRQLtdc	
Administrative censoring at 546 days	admin	

Abbreviations: ECOG, Eastern Cooperative Oncology Group; HRQOL, health-related quality of life
* The time-varying measurements are denoted with 'tdc'

Table C7. Illustration of data setup for a single patient

id	time	trtrand	deathtime	deathtdo	deathind	progtime	xotime	xotdoipcw	disconexp	HRQLtdc
19	0	0	308	0	1	84	84	0	252	.4497
19	21	0	308	0	1	84	84	0	252	.4326
19	42	0	308	0	1	84	84	0	252	.5478
19	63	0	308	0	1	84	84	0	252	.4914
19	84	0	308	0	1	84	84	1	252	.4543
19	105	0	308	0	1	84	84	.	252	.3072
19	126	0	308	0	1	84	84	.	252	.5054
19	147	0	308	0	1	84	84	.	252	.6007
19	168	0	308	0	1	84	84	.	252	.6174
19	189	0	308	0	1	84	84	.	252	.5086
19	210	0	308	0	1	84	84	.	252	.3146
19	231	0	308	0	1	84	84	.	252	.4921
19	252	0	308	0	1	84	84	.	252	.6302
19	273	0	308	0	1	84	84	.	252	.4474

19	294	0	308	1	1	84	84	.	252	.5461

Stata code for the base-case analyses presented in the appendix is provided below.

```
************
*** IPCW ***
************

use "case_study.dta", clear

rcsgen time, df(4) if(xotdoipcw==1) gen(timexosp) //construct spline function of
time for
the WD models

logistic xotdoipcw PROGNOSISb CEAb ECOGb HRQLb timexosp* if trtrand==0 //estimate
probability of switching using baseline predictors only (check model fit)

predict pxo1 if e(sample)

logistic xotdoipcw PROGNOSISb CEAb ECOGb HRQLb PROGTYPEtdc CEAtdc ECOGtdc HRQLtdc
timexosp* if trtrand==0 & progtdc>0 // estimate probability of switching using
baseline and time varying predictors (check model fit)

predict pxo2 if e(sample)

replace pxo2 = 0 if trtrand==0 & !e(sample)
sort id time
gen num = 1-pxo1 if firstobs
replace num = num[_n-1] * (1-pxo1) if !firstobs
gen denom = 1-pxo2 if firstobs
replace denom = denom[_n-1] * (1-pxo2) if !firstobs
gen weight = 1 / denom if trtrand==0
gen sweight = num / denom if trtrand==0 // now explore the distribution of weights

replace weight = 1 if trtrand==1 // set weights to 1 in the treatment arm
replace sweight = 1 if trtrand==1

rcsgen time, df(4) gen(timesp) //construct spline function of time for the FO model

logistic deathtdo trtrand PROGNOSISb CEAb ECOGb HRQLb timesp* [pw=weight] if
xotdoipcw==0, cluster(id) // analysis using unstabilized weights

logistic deathtdo trtrand PROGNOSISb CEAb ECOGb HRQLb timesp* [pw=sweight] if
xotdoipcw==0, cluster(id) // analysis using stabilized weights


**************
*** RPSFTM ***
**************

use "case_study.dta", clear

keep if time==0
stset deathtime deathind, id(id)
replace xotime=deathtime if xoind==0

strbee trtrand, xo0(xotime xoind) endstudy(admin) hr test(logrank) graph gen(t0)

stset t0 dt0, id(id)

sts graph, by(trtrand) // compare counterfactual survival times across randomized
groups


***************************
*** Two-stage adjustment ***
```

```
**************************

use "case_study.dta", clear

drop if trtrand==1 | progind==0 | time<progtime
sort id time
gen tong = 0
by id: replace tong = 1 if (time>=xotime) //create time-varying covariate
indicating a switch to the experimental treatment

replace deathtime = (deathtime - progtime)
replace time = (time - progtime) // changing time variables to indicate time since
disease progression rather than time since randomisation

by id: gen HRQLprog = HRQLtdc[1]
by id: gen ECOGprog = ECOGtdc[1] // use values of HRQL and ECOG at disease
progression

sort id time
by id: gen finalobs = 0
by id: replace finalobs = 1 if _n==_N
expand 2 if finalobs==1
drop finalobs
sort id time
by id: gen finalobs = 0
by id: replace finalobs = 1 if _n==_N
by id: replace time = deathtime if finalobs==1
by id: drop if time[_n+1]==time
by id: gen deathindtd = 0
by id: replace deathindtd = 1 if deathind==1 & finalobs==1
stset time, failure(deathindtd) id(id)
streg tong PROGNOSISb CEAprog HRQLprog ECOGprog progtime, dist(weibull) time //
accelerated failure time model to calculate effect of treatment switching on
survival (check model fit)
di exp(-_b[tong])
scalar accel=exp(-_b[tong]) // save acceleration factor

use "case_study.dta", clear
sort id time
collapse (max) trtrand time deathtime xotime deathind disconexp admin progtime
xoind PROGNOSISb, by(id)
replace xotime=0 if xotime==.
gen tpxo=(deathtime-xotime)
gen txot=xotime
gen timecf = deathtime
replace timecf = (xotime + (tpxo*accel)) if trtrand==0 & xotime>0
gen trecens = timecf //now re-censoring survival times, using 'admin' time to
indicate the end study time for each patient
replace trecens =(admin*accel) if (trtrand==0)
replace deathind = 0 if (timecf>trecens & trtrand==0)
replace timecf = trecens if (timecf>trecens & trtrand==0)

stset timecf, failure(deathind) id(id)
stcox trtrand

*For calculating a confidence interval around the switching-adjusted treatment
effect it necessary to bootstrap the entire estimation procedure
```

APPENDIX REFERENCES

1. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials--an economic evaluation context: methods, limitations, and recommendations. Medical Decision Making. 2014;34(3):387-402.
2. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting for treatment switching in randomised controlled trials - A simulation study and a simplified two-stage method. Statistical Methods in Medical Research. 2017;26(2):724-51.
3. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Dore C, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. Journal of the American Medical Association. 2017;318(23):2337-43.
4. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology (Cambridge, Mass). 2000;11(5):561-70.
5. Watkins C, Huang X, Latimer N, Tang Y, Wright EJ. Adjusting overall survival for treatment switches: commonly used methods and practical application. Pharmaceutical Statistics. 2013;12(6):348-57.
6. Ouwens M, Hauch O, Franzen S. A Validation Study of the Rank-Preserving Structural Failure Time Model: Confidence Intervals and Unique, Multiple, and Erroneous Solutions. Medical Decision Making. 2018;38(4):509-19.
7. Latimer NR, White IR, Abrams KR, Siebert U. Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times? Statistical Methods in Medical Research. 2018.
8. Latimer NR, Henshall C, Siebert U, Bell H. Treatment switching: statistical and decision-making challenges and approahces. International Journal of Technology Assessment in Health Care. 2016;32(3):160-6.

9. White IR, Walker S, Babiker A. strbee: Randomization-based efficacy estimator. Stata Journal. 2002;2(2):140-50.
10. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology (Cambridge, Mass). 2011;22(1):42-52.