Article:
Kapetanakis, V., Prawitz, T., Schlichting, M. et al. (2019) Assessment-schedule matching in unanchored indirect treatment comparisons of progression-free survival in cancer studies. PharmacoEconomics, 37 (12). pp. 1537-1551. ISSN: 1170-7690

https://doi.org/10.1007/s40273-019-00831-3

Title:	**Assessment-schedule matching in unanchored indirect treatment comparisons of progression-free survival in cancer studies.**
Authors:	Venediktos Kapetanakis, Thibaud Prawitz, Michael Schlichting, Jack Ishak, Hemant Phatak, Mairead Kearney, John W Stevens, Agnes Benedict, Murtuza Bharmal
Journal:	PharmacoEconomics
Contact information:	venediktos.kapetanakis@evidera.com , Tel: +44 20 8576 5058
Key words:	Evaluation time bias, interval-censoring, health technology assessment, relative efficacy, bias correction, assessment time bias
Funding:	This research was funded by Merck Healthcare KGaA, Darmstadt, Germany, and is part of an alliance between Merck Healthcare KGaA and Pfizer Inc, New York, NY, USA

Abstract

Background

The timings of efficacy-related clinical events recorded at scheduled study visits in clinical trials are interval-censored with interval duration pre-determined by study protocol. Events may happen anytime during that interval but can only be detected during a planned or unplanned visit. Disease progression in oncology is a notable example where time-to-event is affected by the schedule of visits within a study. This can become a source of bias when studies with varying assessment-schedules are used in unanchored comparisons using methods such as matching-adjusted indirect comparisons.

Objective

We illustrate assessment-time bias (ATB) in a simulation study based on data from a recent study in second-line treatment for locally advanced or metastatic urothelial carcinoma,and present a method to adjust for differences in assessment schedule when comparing progression-free survival against a competing treatment.

Methods

A multi-state model for death and progression was used to generate simulated death and progression times from which PFS times was derived. PFS data were also generated for a hypothetical comparator treatment by applying a constant HR to the baseline treatment. Simulated PFS times for the two treatments were then aligned to different assessment schedules so that progression events were only observed at set visit times, and the data were analysed to assess the bias and standard error of estimates of hazard ratios between two treatments with and without assessment-schedule matching (ASM).

Results

ATB is highly affected by the rate of the event at first assessment time; in our examples, the bias ranged from 3% to 11% as the event rate increased. The proposed method relies on individual-level data from a study and attempts to adjust the timing of progression events to the comparator's schedule by shifting them forward or backward without altering the patients' actual follow-up time. The method removed the bias almost completely in all scenarios without affecting the precision of estimates of comparative effectiveness.

Conclusions

Considering the increasing use of unanchored comparative analyses for novel cancer treatments based on single-arm studies, the proposed method offers a relatively simple means of improving the accuracy of relative benefits of treatments on progression times.

Key Points for Decision Makers

1. Indirect treatment comparisons based on single-arm studies may be biased by differences in schedules for tumour imaging, favouring the treatment with visits scheduled with longer intervals.
2. The size of the bias is highly affected by the distribution of events over time and becomes larger when these occur early, in which case even relatively small differences in schedule (e.g. 2 weeks) may be sufficient to bias estimates of relative effect and generate misleading conclusions
3. Assessment-schedule matching offers a relatively simple and adaptable means of adjusting for this bias and should be used prior to conducting unanchored comparisons of treatments in health technology assessments such as population-adjusted indirect comparisons.

1. Introduction

Time-to-event outcomes are often measured on a continuous scale based on the time elapsed between some meaningful starting point (e.g., date of randomisation or start of treatment) until the date at which the event of interest (e.g., death) occurs. Some types of events such as response-to-treatment or progression of disease are evaluated at scheduled study visits and cannot be measured at the exact time they occur. The time at which the event is recorded is interval-censored, meaning that it must have occurred sometime between the current and last visit when the patient was known to be event-free.

A notable example is progression-free survival (PFS) in studies of cancer treatments. PFS is a composite endpoint defined as the earlier of death or progression of disease, with those patients not experiencing either event treated as censored at the end of follow-up. While exact dates of death are usually known, progressions are mostly identified at assessments conducted at pre-determined intervals as per study protocol. Scheduled assessments can vary according to the type of treatment, and hence, across different studies. Furthermore, in practice the exact schedule may deviate from the planned schedule.

Interval-censored outcomes are commonly analysed by treating events recorded at assessment times as exact. However, the nature of interval-censored outcomes implies that estimates of the risk of the outcome are underestimated at times between scheduled visits, as events are recorded with a systematic delay at the following visit. Estimation of progression-free survival can be biased if intermittent assessment of progression is not appropriately considered.[1] Panageas et al.[2] conducted a simulation study demonstrating this bias according to various definitions of PFS, but found that comparative analyses between groups following the same assessment schedule were not affected. Any distortions associated with the schedule are common to the groups and likely cancel out in estimates of relative effect such as hazard ratios (HRs). Similar findings were reported by Qi et al. in a later study.[3]

While these studies would suggest that estimates of relative treatment effect within studies may be robust, the same cannot be assumed for comparisons of outcomes with treatments from different studies following different schedules. This may occur, for example, when unanchored indirect treatment comparisons (ITCs) are either based on single-arm studies or, in the absence of a common reference arm, on randomized trials. These ITCs are becoming increasingly common in health technology assessments (HTAs), especially for novel cancer treatments approved using single-arm trials conducted in high unmet need population. Between 2009 and 2014, the US Food and Drug Administration (FDA) assessed 54 drugs in 64 indications based on single-arm studies. The European Medicine Agency (EMA) assessed 35 drugs in 44 indications over the same period.[4] As of November 2017, Alexiou et al.[5] identified 13 National Institute for Health and Care Excellence (NICE) technology appraisals supported by single-arm studies.

Unanchored ITCs based on single-arm studies are prone to several potential challenges including potential confounding bias because of differences in the populations of the studies with respect to treatment effect modifiers and prognostic factors of the outcome.[6] Methods such as matching-adjusted indirect comparisons (MAIC)[7] or simulated treatment comparisons (STC)[8] can address such imbalances, but cannot address bias arising from different assessment schedules used in the studies. For example, suppose two treatments are studied in single-arm studies

with identical populations and have identical clinical efficacy (i.e., risk of progression), but in one study disease progression is assessed every 4 weeks whereas it is assessed every 9 weeks in the other study. The treatment in the second study will falsely appear to have a lower risk of progression than the treatment in the first study and will lead to biased estimates of relative effect.

We define this phenomenon as assessment-time bias (ATB) and propose a method to reduce the bias in estimates of relative effect on PFS in ITCs of single-arm studies with differing assessment schedules. The described method for assessment-schedule matching (ASM) is shown to correct the bias in a simulation study varying some of the key parameters affecting the size of bias.

2. Context of the Problem

The discussion of ATB and the proposed ASM method will be framed in the context of an unanchored ITC of PFS (i.e., when there is a disconnected treatment network or single-arm studies). It is assumed that individual patient data (IPD) are available for one of the studies; we refer to this as the index study and index treatment. Data for the second or comparator study are assumed to only be available from a publication. A MAIC or STC in this context would first estimate an adjusted PFS survival function for the index treatment which reflects the expected survival in the comparator's population. This is then contrasted with the comparator treatment's PFS survival function to estimate a relative effect such as a HR. We postulate that when the schedule of assessment of disease progression differs between two studies, ASM is necessary prior to deriving the estimate of relative effect. This paper will focus specifically on this correction step, as other steps of the comparative analyses would be applied per usual.

We use a case study based on an immunotherapy study for bladder cancer to illustrate the method. PFS survival functions for these treatments are characterized by a sharp drop at the first assessment time followed by a relatively steady decline and, in some cases, an eventual plateau. It is not uncommon for a significant portion (often more than half) of the patients to have experienced disease progression or died prior to the first assessment time. Assessment schedules for studies will determine timing of assessments to capture the median by this point. Figure 1 shows representative examples of PFS survival functions estimated in recent immunotherapy studies in advanced urothelial and gastric cancer.[9-12] In all cases, the survival functions drop below the median at their respective first assessment time; assessment time is, therefore, likely to be the most important driver of bias because most of the events observed in the study have already occurred. It may be sufficient for studies to differ only in the first assessment time for bias to occur. A study with a longer assessment interval may also be more likely to miss disease progression in patients who drop out of the study early, further slowing the drop in the observed PFS survival function.

3. Assessment-schedule matching (ASM)

3.1 Overview and Notation

The proposed method aims to adjust the PFS survival function of the index treatment to reflect what would have been observed if the study had followed the comparator's assessment schedule for progressive disease. The adjustment must be made in this direction because individual patient-level data are only available from the index study. Furthermore, this is consistent with the direction of adjustment in the MAIC or STC for which the correction is required.

A PFS event can occur in one of three ways: a patient dies prior to disease progression being observed; disease progression is recorded at an unplanned visit between scheduled study assessments; or disease progression is recorded at a scheduled assessment time. Patients not experiencing one of these three scenarios are censored for PFS, typically on the date of their last tumour assessment. The method only adjusts events recorded at a scheduled assessment time. Death times are known exactly and are unaffected by the studies' assessment schedules; the same can be assumed for disease progression observed at unplanned medical visits because these are prompted by aggravation of the patients' condition. Thus, the timing and occurrence of these events are not subject to adjustment.

Suppose that assessments are made at times $T_1, T_2, T_3, \ldots, T_K$ in the index study, while the comparator study scheduled visits are at $T_1^*, T_2^*, T_3^*, \ldots, T_M^*$. It is possible that assessment times are not perfectly interspersed; that is, we may not necessarily have $T_1 < T_1^* < T_2 < T_2^* < T_3 < T_3^*, \ldots$. Indeed, studies may only differ in the first few visits and subsequently align; for example, theoretically the index study may schedule visits at 4, 8, 12 weeks, and every 12 weeks thereafter, and the comparator study conducts assessments at weeks 6, 12 and every 12 weeks thereafter. It is also possible that multiple visits from one study fall in the interval of the other study's schedule. For instance, we may have $T_1 < T_2 < T_1^* < T_3 < T_2^*, \ldots$. To simplify the discussion, and without loss of generality, we assume that in these cases, events from the multiple visits falling between visits scheduled in the comparator study can be pooled and processed together, so that we can conceptualise these as a single visit. In what follows we assume that at most one index visit happens between comparator visits, and that the comparator study follows a longer interval at least at the first assessment i.e., $T_1 < T_1^* \leq T_2 < T_2^* \leq T_3 < T_3^*, \ldots, \ldots$ We consider how to handle cases following a different pattern further below.

The method proceeds by estimating how many progressions from the index study would be captured at each T_i^* by shifting the events captured at visit T_i forward, and an appropriate proportion (p) of those captured at T_{i+1} backward. Additionally, the method ensures that any shifted progression times do not exceed death or censoring times, as these progressions would have been missed at T_i^*. For instance, suppose 20 patients have progressed at $T_1 = 6$ weeks, one of whom then dropped out of the study at week 7. If the first assessment had been scheduled at $T_1^* = 8$ weeks, only 19 of the 20 progressions would have been captured. Suppose further that 9 patients had progressed at $T_2 = 12$ weeks; some fraction of these would have occurred between 6 and 8 weeks and, therefore, captured at $T_1^* = 8$. A simple approximation would be that this fraction is $p = \frac{T_1^* - T_1}{T_2 - T_1} = \frac{2}{6} = 0.33$, assuming an even distribution of the events in the interval. This implies that 3 of the 9 events would have been picked up at T_1^*, leading to a total of 22 progressions.

Figure 2 illustrates the adjustment process; the details of the approach, and more precise means of deriving p are described in the following sections

3.2 ASM Method

We denote by τ_j the PFS time of the j^{th} patient in the index study, and by ρ_j their status where 1 indicates they progressed or died, and 0 indicates they were censored for PFS, i.e., event-free and alive at the end of their follow-up. We aim to derive τ_j^* and ρ_j^*, the PFS time and status we would have observed if the index study followed the assessment schedule of the comparator study.

As noted above, not all progression events in the index study are subject to adjustment. Patients who had death as the first event recorded or a progression detected at an unplanned visit between scheduled assessments would be expected to have these same times recorded under the comparator's schedule. Thus, for these patients $\tau_j^* = \tau_j$ and $\rho_j^* = \rho_j$. The same is assumed for censoring times. This is a conservative approach for the index treatment because forward shifting of censoring times would require assuming that the patient was followed longer than observed; adjustment by backward shifting would not be applicable because the patient is known to be progression-free up to the time of censoring so any evaluation of disease progression before censoring would capture the progression-free status of the patient and this would be captured by the proposed algorithm.

Adjustment is applied only for patients who had disease progression recorded at a scheduled assessment time. For these patients, the progression time will correspond to one of $T_1, T_2, T_3, \ldots, T_K$, but may vary slightly because not all visits would fall exactly on schedule to the day. We denote by n_i the number of patients whose disease progression was recorded at T_i.

The adjustment process then iterates through each of the assessment schedules of the comparator study (T_i^*), and shifts progressions from visits immediately preceding and following this time in the index study, where it is assumed that $T_1 < T_1^* \leq T_2 < T_2^* \leq T_3 < T_3^*, \ldots$.

3.2.1 Adjustment of the First Visit

We first describe the steps involved in adjusting the progression times and status for patients whose disease progression was recorded at T_1.

Step 1. Forward Shift

The first step involves advancing the times of individual disease progression at T_1 in the index study to T_1^*, because these events would only have been detected at that time. This is done based on individual progression times to preserve the observed variation around visit dates, and only in cases where these occur at least one week earlier than T_1^* (i.e., $T_1^* - T_{1j} > 1$week) to reflect the fact that studies typically allow a buffer around scheduled assessments. Thus, those events already falling within a week of the comparator's assessment time may have counted as the patient's study visit and not shifted further.

The adjusted time for these patients is then given by

$$\tau_j^* = \begin{cases} T_{1j} + (T_1^* - T_1) & \text{if } T_1^* - T_{1j} > 1 \\ T_{1j} & \text{if } T_1^* - T_{1j} \leq 1 \end{cases}$$

for $j = 1, \dots, n_1$.

Step 2. Correction for Death or Censoring

Death or censoring time, whichever occurs first, and status are denoted $D_j$and δ_j (1 if the patient dies and 0 if censored) and are assumed fixed and not subject to adjustment. Thus, if τ_j^* exceeds D_j, it is assumed that the progression event would be missed at T_1^*. The adjusted PFS time is then set to $\tau_j^* = D_j$, and $\rho_j^* = \delta_j$. Thus, a patient who was censored at D_j, would no longer be counted as an event; a patient who had died would count as an event at the time of the death rather than the adjusted progression time.

Step 3. Backward Shift

The final step of the algorithm involves shifting events from T_2 back to T_1^*, since some of these events would be expected to occur between T_1 and T_1^*. Three approaches are considered to estimate the proportion of events to be shifted backward (p_1).

Approach 1: Linear Interpolation

A simple way to estimate $p_1$is by linear interpolation; that is, assuming events in the period occur uniformly, and therefore, the proportion $p_1 = \frac{T_1^* - T_1}{T_2 - T_1}$ of the events would have been captured at T_1^*.

Approach 2: Progression Probability-Based Calculation

To capture variation in risk of disease progression in the interval, p_1 may be calculated based on the proportional change in the survival probabilities at T_1, $T_1^*$and T_2. That is:

$$p_1 = \frac{S(T_1) - S(T_1^*)}{S(T_1) - S(T_2)},$$

where $S(\cdot)$ are probabilities of being progression-free derived from a parametric survival model fitted for time to progression (TTP) while accounting for the interval-censored nature of progression times.

Approach 3: Worst-case Scenario

Alternatively, p_1 can be set to 1 assuming that all n_2 patients would have progressed by T_1^*. This is likely an unrealistic scenario; however, there is value in exploring this approach as a "worst-case scenario".

For a given p_1, the progression times are shifted back to T_1^* (i.e. $\tau_j^* = T_1^*$) for $n_2^* = \text{ceiling}(p_1 \times n_2)$ out of the n_2 patients observed to progress at T_2, where $\text{ceiling}(x)$ is the smallest integer greater than or equal to x.

Figure 3 illustrates the adjustment steps at the first visit.

3.2.2 Adjustment of Subsequent Visits

The process for adjustment of subsequent visits follows the same steps as described above, with one change to step 1. Due to the backward shift in step 3, only a portion of the patients progressing at later visits are available to be shifted forward. Thus, for patients progressing at visits T_i for $i \geq 2$, step 1 would derive adjusted times as follows:

$$\tau_j^* = \begin{cases} T_{ij} + (T_i^* - T_i) & \text{if } T_{ij} - T_i^* > 1 \\ T_{ij} & \text{if } T_{ij} - T_i^* \leq 1 \end{cases}$$

where j references the set of $n_i - n_i^*$ patients that were not shifted backwards at the previous iteration of the algorithm. Steps 2 and 3 would follow as above.

It is worth noting that in situations where the visits in the studies coincide (i.e., $T_i^* = T_i$), the steps of the algorithm will inherently produce no change in the observed event times as the adjustment $(T_i^* - T_i)$ would equate to 0.

The adjustment can proceed iteratively up to the latest assessment time between the two studies (i.e., $\min(T_K, T_M^*)$). If $T_K > T_M^*$, an adjustment would not be possible at visit T_K, and the process stops after adjusting T_{K-1}; if, on the other hand, $T_K < T_M^*$, step 3 cannot be implemented for T_K, and the process stops after step 2.

3.2.3 Dealing with Other Assessment Patterns

The discussion so far has assumed $T_1 < T_1^* \leq T_2 \leq T_2^* \leq T_3 \leq T_3^*$, ... In practice, any number of patterns, both regular and irregular, may be observed. One specific scenario of interest is when the assessment intervals are longer in the index study: $T_1^* < T_1 \leq T_2^* \leq T_2 \leq T_3^* \leq T_3$ .... Given that adjustment is done based on the index study where IPD are available, forward shift is not possible as the first step in the algorithm. In this scenario, progressions at T_i must first be shifted backward (step 3) based on the appropriate proportion, p_i; the balance of the patients are then shifted forward as described in steps 1 and 2. When the intervals follow an irregular pattern where there is no consistent ordering between the index and comparator studies, the algorithm must consider the sequence and apply the steps accordingly.

3.2.4 Adjustment for All vs. Some Visits

While it is possible to adjust all visits, it is advisable to consider whether this is necessary or even beneficial. For instance, in the context of PFS survival functions that are characterised by a sharp drop at the first assessment time, ATB is likely to be driven mostly by the misalignment of visits at that first occasion. Thus, it may be possible to remove potentially ATB completely by applying adjustment at this point. In addition to simplicity, this has the added advantage of minimizing alterations to the index data, particularly the re-censoring of some of the progressions in step 2. A third advantage in applying ASM at the first visit only is that reliance on the interval-censored model in step 3 is minimised as this model may provide less reliable estimates at later times because of progressive loss of power particularly when data are immature. A focused adjustment (i.e. on first assessment visit) would also be sufficient in cases when the relative schedules have an irregular order where distortions may naturally cancel out.

The assessment of the statistical properties of the approach using simulations considered both the scenario of adjusting the first assessment time only as a primary approach and the added value of full adjustment.

4. Simulation Study

4.1 Overview

To illustrate and assess the statistical properties of the method, we conducted a simulation study based on the PFS survival function estimated in a recent study of patients receiving second line treatment for locally advanced or metastatic urothelial carcinoma.[13] This required first developing a model for PFS based on the study data. A multi-state model for death and progression was used and calibrated to capture the early rate of progression (i.e., sharp drop in the PFS survival function) accurately. The model was then used to generate simulated death and progression times from which PFS times can be derived under different scenarios, varying parameters like the rate of early

progression, sample size and duration of follow-up. PFS data were also generated for a hypothetical comparator treatment by applying a constant HR to the baseline treatment. Simulated PFS times for the two treatments were then aligned to different assessment schedules so that events were only observed at set visit times. The schedules used for the two treatment groups were varied in the simulation, as was the approach used to derive the backward shift proportion. Simulated data were analysed to assess the bias and standard error of estimates of HRs between treatments with and without ASM.

4.2 Modelling PFS for Simulation

In the study used as a basis for the simulation, PFS was measured by radiographic imaging following a six-weekly schedule for the first 12 months, and a twelve-weekly schedule thereafter. Figure 4 shows the empirical PFS survival function used as a basis to simulate data to test the approach. This follows the expected pattern with nearly half of patients having an event by the time of the first scheduled assessment.

A model was needed from which actual PFS times could be simulated on a continuous scale (i.e., independent of any scheduled visit times). A piecewise multi-state model (Figure 5) based on an exponential distribution over weeks 0 to 6 (the first assessment time), and a Weibull distribution thereafter was used. This approach was selected to enable the differentiation of PFS events into progressive disease and deaths and to facilitate the investigation of different scenarios on the magnitude of the initial decline in PFS at 6 weeks. Furthermore, to reflect the PFS at 6 weeks accurately, the parameter of the exponential distributions was calculated based on the observed survival proportions for each of the three outcomes. The Weibull segments were estimated by maximum likelihood. This model provided a good fit to the data (Figure 4, green curve) and was adopted for the simulation.

PFS times were simulated for individual patients by inverting the survival functions corresponding to each of the transitions in the model. This produced a triplet consisting of a sampled time for progressing prior to death (PD), dying prior to progression (TTDBP) and dying after progression (PPS). To mimic censoring that would occur in studies, a maximum follow-up (MFU) duration was assumed, and events occurring after this time were considered censored. This is a simplification of what happens in studies where patients may have variable follow-up duration because of staggered enrolment or early drop out, but we do not expect this to affect our analyses because censored values are not modified in the method.

If the sampled PD or TTDBP values did not exceed the maximum follow-up time, the patient was considered to have a PFS event with $\tau_j = \min(PD_j, TTDBP_j, MFU)$ and $\rho_j = 1$ if event times preceded MFU, and 0 otherwise. PPS times were used to calculate overall survival or censoring time: $D_j = \min(PD_j + PPS_j, MFU)$ and event status (δ_j), which are needed in the adjustment process. PFS times were also simulated for a hypothetical comparator study following this process after applying a fixed HR to each of the hazard functions in the model.

To induce interval-censoring, the index and comparator PFS times were aligned to two different assessment schedules. All progression events were moved to, and the PFS times were reset to the closest assessment time following the event (e.g., 6 weeks). No stochasticity was applied to create distortions in individual visits around the assessment time. This may underestimate the variability that occurs in actual studies but is not expected to bias our analyses in a systematic way.

The simulation process was replicated to generate 1000 datasets under different scenarios, each containing observed and ATB-adjusted PFS times for the index group and observed times for the comparator. Cox proportional hazards models were fitted to obtain estimates of the HR between treatments based on the observed and adjusted times after ASM. These estimates were contrasted with the true value during simulation to quantify the bias. Mean estimate of HRs across replications, percentage bias, 95% confidence interval limits, coverage probabilities (i.e. the proportion of replications when the estimated confidence interval contained the true HR), root mean square error (rMSE) and mean standard error of log HR were derived.

4.3 Simulation Parameters and Scenarios

The parameters varied in the simulation are summarized in

Table 1. In addition to sample size and HR, PFS at the first assessment were varied to capture different degrees of potential bias. Different assessment schedules were also considered to assess whether the relative difference in timing affects performance. Finally, three different options were considered for derivation of the backward shift proportion (p_i) using the approaches described above. Numbers in bold apply to the base case scenario simulation settings; alternative scenarios were then investigated by varying each parameter one at a time. In all cases, ASM was applied only at the first assessment point; full adjustment was assessed in the every 6 vs. 9 weeks and every 6 vs. 8 weeks scenarios as the difference in schedules in other cases were relatively small.

5. Findings from Simulation Study

5.1 Illustration of the Bias

The presence and potential extent of bias according to varying assessment schedules is shown in Figure 6. Specifically,

Figure 6 shows actual and simulated PFS survival functions under differing assessment schedules and varying the observed PFS at the first assessment. The simulated survival functions are generated from the same underlying model, assuming a HR of 1.0. To illustrate the bias, we recorded the median time observed with each assessment schedule,and quantified the overall difference between the survival functions in terms of a hazard ratio. In the absence of bias, the medians would be equal in the two scenarios and the hazard ratio would equal 1. Results are summarized in Table 2 and show that the bias at the median of these curves increases with the proportion of events observed prior to the first scheduled visit and is larger when assessments are scheduled later. For instance, when the PFS at the first assessment is 0.4, the median from the comparator schedule was estimated to be almost double of the actual value. Similarly, HRs indicate an overall relative difference of 11% between the survival functions when 40% were progression-free at the first schedule, and 3% when PFS was 80% at the first visit.

5.2 Results from Simulated Scenarios

Results of all scenarios investigated are presented in Table 3. In the base case scenario where the true effect was a HR of 1.25, a crude comparison of the index and comparator PFS times led to HRs that were consistently underestimated by 8%, on average. Figure 7 shows the empirical distribution of the estimated HRs in replications of the base case scenario with and without ASM. Crude analyses tended to underestimate the true HR. The size of the bias was generally consistent when sample size, HR, duration of follow-up and assessment schedule were varied, but was directly dependent on rate of progression before the first assessment. The bias was -11.3% on average when around 40% were event-free at the first visit, and almost negligible (-2.3%) when around 80% were event-free. Coverage probabilities of the 95% confidence interval were consistently below the nominal values, except in the case where the event-free rate was high.

ASM at the first visit almost completely removed the bias observed in the crude analyses across all scenarios. While very small (generally between 0.1% and 0.6% when varying input parameters), the direction of the bias remained generally positive. This may be explained by the upward rounding of the ceiling function in step 3 of the algorithm. The adjustment was also effective with the linear interpolation (with a bias of -1.0% on average) and worst-case scenario (with a bias of -1.9% on average) approaches to calculating the backward shift proportion, but these were not as effective as those using the probability-based calculations. Coverage probabilities were also restored following ASM, and in one scenario (when sample sizes were 250), the adjustment led to a confidence interval that excluded the null, identifying an effect that would have been missed in the crude analysis. The rMSE was reduced slightly after ASM, driven mostly by the removal of bias in the estimates. The SE of the effect estimates were not affected.

The base case considered schedules that eventually overlap perfectly. When considering schedules that differ more frequently (e.g., every 6 vs 8, or every 6 vs. 9 weeks), ASM at the first visit was sufficient to remove almost all the bias observed in crude comparisons. This was particularly so with a 3-week difference in schedules where the bias was only 0.1% on average. In sensitivity analyses applying correction at all visits, bias remained negligible (<1%) but was found to increase, reflecting that correction at all visits does not produce further improvement in accuracy.

6. Discussion

Indirect treatment comparisons based on single-arm studies may be biased by differences in outcome assessment times, favouring the treatment with visits scheduled with longer intervals. We showed that the size of the bias is highly affected by the distribution of events over time and is particularly important when these occur early – at the first assessment in our example. In these cases, even relatively small differences in schedule (2 weeks in our example) may be sufficient to bias estimates of relative effect and generate potentially misleading conclusions. The proposed ASM method reduced the bias almost completely across all scenarios we assessed. Given that most events in our assessments occurred prior to the first assessment time, adjustment at this point was sufficient and further gains were not observed by adjusting later visits. However, this may not be the case in all situations.

Our example focused specifically on the context of PFS assessment in studies where a large proportion of events occur prior to the first visit. This is common across different cancer types and treatments (Figure 1), and evidence

suggests that the problem might be more pronounced in studies of immunotherapies because some patients may experience hyper-progression causing a sharper drop in estimated survival function.[14] In other applications, or in cases of less aggressive disease, the early drop in the survival function may be less pronounced or potentially spread over the first few assessments rather than only at the first assessment. The potential for bias may be minimal in cases where the survival functions decline steadily. Given that our simulations did not explore such patterns directly, we cannot affirm that adjustment of the first visit is sufficient in all cases. It is, however, advisable to consider ASM beyond the first visit at least in a sensitivity analysis to confirm that the results are robust. While adjusting all visits may be the strictest approach, it requires more manipulation of the original index data, and potential censoring of some observed progressions in step 2 of the algorithm. Furthermore, probabilities used in step 3 would have to be read from the tail of the modelled TTP curve and may not be as reliable as those from earlier portions because of immaturity of data or poor fit of the interval censored model in the tail of the survival function.

The proposed ASM method assumes death and disease progression detected at unplanned visits (while not common in a clinical trial setting) should not be adjusted because these are unlikely to be affected by the visit schedule; furthermore, adjustment of censoring times should not be made to avoid making unverifiable assumptions that could favour the index study (e.g., extending the patients observed duration of follow-up and treating them as progression-free during this period). The method only adjusts progression events that are captured at scheduled assessment visits. The method further assumes that these events would have been only captured at the next scheduled time of the comparator study. It is possible that some of these progression events would have been captured at unplanned visits prior to the comparator's next scheduled visit and captured earlier. However, incorporating adjustments for such unlikely events would add further complexity, with potentially minimal gain in bias reduction considering the results of our simulation. It should also be acknowledged that the proposed method does not revise times-to-event for patients who might have progressed between their last progression assessment and death if their last assessment were to take place later within that interval. Although this may favour the index treatment because the time-to-event in the original data is overestimated, the same limitation applies to an analysis without ASM.

Another assumption in the method is that the estimates of the proportion of events to shift backward in step 3 are accurate. These are based on a parametric model fitted to the observed progression times while accounting for interval-censoring. Their accuracy depends on the goodness-of-fit of the chosen model. In our simulations, the bias correction with this approach was only slightly better than corrections based on linear interpolation, which assume that events occur evenly in the interval and allocates these proportional to time between visits. Thus, the method may be relatively insensitive to misspecification of the progression models, and the added complexity of using probability-based adjustment factors, while more robust, may not always be warranted.

A few simplifications were made in the simulation process. Studies were assumed to have similar duration of follow-up and censoring only occurred at end of study (and hence, no possibility of informative censoring due to selective drop-out affecting results). Furthermore, no unplanned visits were created, and duration of follow-up was capped based on an assumed maximum study duration. Given that the method does not alter these aspects of the data, these simplifications are likely to be inconsequential. Backward shift proportions were also derived once and applied across replications to minimize run-time of the analyses. It is possible that incorporating uncertainty in these proportions across simulations would induce more variability in results across replications, but we expect this to occur in a random way rather than affecting the direction of bias of the estimates. The simulation process assumed that hazards between treatment arms are proportional. Although the proposed method for ASM does not rely on this assumption, its properties have not been investigated in the setting where the assumption of proportional hazards is violated, and other measures of effect are used for the comparison (e.g., an acceleration or shrinkage factor in an accelerated failure time model, differences in means or medians, etc.).

It is important to note that ASM only corrects for the scheduling between studies being compared. Indirect treatment comparisons of these would have to further consider adjustment for prognostic factors and treatment effect modifiers using methods such as MAIC or STC. It is also important to consider that PFS times remain interval-censored after adjustment. These approaches can be applied following their usual steps after applying ASM to the index study.

While the problem of progression times being interval-censored is understood, there are few options to adjust for these analytically. Tanase et al. [15] proposed an exploratory analysis to re-analyse PFS by shifting progression assessment schedule. However, we were not able to compare our method to theirs because they did not provide sufficient information about their approach and underlying assumptions. Other strategies may be possible, such as those proposed by Panageas et al.[2] and Heller et al.[16], to deal with interval-censoring of PFS. While appealing, these methods require patient-level data from both studies being compared, which is typically not possible in the context addressed here.

Although the paper has focused on applications in cancer studies and PFS specifically, the method can be adapted to any situation where events are assessed on a discrete schedule. The focus on PFS is relevant and important given the clinical landscape with increasingly more treatments being subjected to review while still at early phases of development, particularly with novel treatments such as immunotherapies. When evidence comes from single-arm studies or disconnected networks, decisions about the use of such treatments must rely on comparative effectiveness analyses using methods for unanchored ITC such as MAIC or STC. Waiting for evidence from randomized controlled trials and anchored comparisons (e.g. network meta-analysis) may not be possible because of ethical considerations. In addition to the commonly known challenges associated with unanchored comparisons (e.g., residual confounding), our paper shows that ATB may be another important and substantial source of distortion of results that should be considered when performing such comparisons.

7. Conclusion

In summary, the proposed method to correct for ATB offers a relatively simple and adaptable means of adjusting for these distortions in unanchored comparisons of treatments that may be important when conducting a health technology assessment.

8. Data Availability Statement

The software code for the generation of the data analysed during this study and the implementation of the assessment-schedule matching method is included in the supplementary information files of this published article.

9. Compliance with Ethical Standards

9.1 Funding

This research was funded by Merck Healthcare KGaA, Darmstadt, Germany, and is part of an alliance between Merck Healthcare KGaA and Pfizer Inc, New York, NY, USA.

9.2 Conflict of interest

Venediktos Kapetanakis, Thibaud Prawitz, Jack Ishak and Agnes Benedict are employees of Evidera that was hired by the sponsor, Merck Healthcare KGaA to conduct this research. John Stevens served as a consultant to Evidera. Michael Schlichting and Mairead Kearny are employees of the sponsor, Merck Healthcare KGaA. Hemant Phatak and Murtuza Bharmal are employees of EMD Serono, a business of Merck KGaA, Darmstadt, Germany.

9.2 Acknowledgments

Venediktos Kapetanakis and Jack Ishak conceived the method. Michael Schlichting, Hemant Phatak, Murtuza Bharmal and John W Stevens contributed to the method inception. Thibaud Prawitz analysed the data and wrote the first draft of the manuscript. Murtuza Bharmal led the project team from inception to completion. Venediktos Kapetanakis, Michael Schlichting, Jack Ishak, Hemand Phatak, Mairead Kearny, John W Stevens, Agnes Benedict and Murtuza Bharmal contributed to the interpretation of results and revision of the manuscript. All authors have read and approved the final manuscript. Venediktos Kapetanakis is the guarantor of the manuscript.

References

1. Zeng L, Cook RJ, Wen L, Boruvka A. Bias in progression-free survival analysis due to intermittent assessment of progression. *Stat Med.* 2015;34(24):3181-3193.
2. Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D. When you look matters: the effect of assessment schedule on progression-free survival. *J Natl Cancer Inst.* 2007;99(6):428-432.
3. Qi Y, Allen Ziegler KL, Hillman SL, et al. Impact of disease progression date determination on progression-free survival estimates in advanced lung cancer. *Cancer.* 2012;118(21):5358-5365.
4. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ Open.* 2016;6(6):e011666.
5. Alexiou D, I. Chatzitheofilou, and A. Pi Blanque. A Review of Nice Technology Appraisals in Oncology Using Single Arm Trials (SAT) Evidence. *Value in Health.* 2018;21:S224.
6. Phillippo DM AA, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU TECHNICAL SUPPORT DOCUMENT 18: METHODS FOR POPULATION-ADJUSTED INDIRECT COMPARISONS IN SUBMISSIONS TO NICE. 2016.
7. Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics.* 2015;33(6):537-549.
8. Signorovitch JE, Wu EQ, Betts KA, et al. Comparative efficacy of nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison of randomized trials. *Curr Med Res Opin.* 2011;27(6):1263-1271.
9. NICE Single technology appraisal ID939: Atezolizumab for treating metastatic urothelial bladder cancer after platinumbased chemotherapy.
10. Fuchs CS, Doi T, Jang RW, et al. Safety and Efficacy of Pembrolizumab Monotherapy in Patients With Previously Treated Advanced Gastric and Gastroesophageal Junction Cancer: Phase 2 Clinical KEYNOTE-059 Trial. *JAMA Oncol.* 2018;4(5):e180013.
11. Kang YK, Boku N, Satoh T, et al. Nivolumab in patients with advanced gastric or gastro-oesophageal junction cancer refractory to, or intolerant of, at least two previous chemotherapy regimens (ONO-4538-12, ATTRACTION-2): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet.* 2017;390(10111):2461-2471.
12. Powles T, O'Donnell PH, Massard C, et al. Efficacy and Safety of Durvalumab in Locally Advanced or Metastatic Urothelial Carcinoma: Updated Results From a Phase 1/2 Open-label Study. *JAMA Oncol.* 2017;3(9):e172411.
13. Apolo AB, Ellerton J, Infante JR, et al. Avelumab treatment of metastatic urothelial carcinoma (mUC) in the phase 1b JAVELIN Solid Tumor study: updated analysis with ≥12 months of follow-up in all patients. 42nd ESMO Annual Congress; 8-12 September 2017, 2017; Madrid, Spain.
14. Champiat S, Dercle L, Ammari S, et al. Hyperprogressive Disease Is a New Pattern of Progression in Cancer Patients Treated by Anti-PD-1/PD-L1. *Clin Cancer Res.* 2017;23(8):1920-1928.
15. Tanase T, Hamada C, Yoshino T, Ohtsu A. A Proposal for Progression-Free Survival Assessment in Patients with Early Progressive Cancer. *Anticancer Res.* 2017;37(10):5851-5855.
16. Heller G. Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime Data Anal.* 2011;17(3):373-385.

Table 1. Parameters that were varied in simulations

Varying parameter	**Parameter values**	
Sample size	75, **100**, 250	
HR (comparator vs index treatment)	0.80 (better PFS for comparator), 1 (no treatment effect), **1.25 (worse PFS for comparator)**	
PFS at T_1	0.4, **0.6**, 0.8	
Maximum follow-up (weeks)	52, **104**, 156	
Progression assessment schedules	Index study	Comparator study
	6, 12 and every 4 weeks afterwards	**8, 16 and every 4 weeks afterwards**
	Every 6 weeks	Every 8 weeks
	Every 6 weeks	Every 9 weeks
Backward shift proportion calculation	**Progression Probability,** Linear approximation, worst-case	

PFS: progression-free survival, T_1: First progression assessment time in index study. Parameters used in the base case scenario are emphasised in bold.

Table 2. Summary of bias when different assessment schedules are imposed on the same underlying PFS curves.

	Progression-free survival at 6 weeks					
	40%		**60%**		**80%**	
	Estimate (95% CI)	**Difference**	**Estimate (95% CI)**	**Difference**	**Estimate (95% CI)**	**Difference**
Median PFS						
True time to event records	4.2 (3.0, 5.3)	-	7.5 (4.2, 10.9)	-	18.3 (9.6, 27.0)	-
Simulated PFS with visits at 6, 12 and every 4 weeks afterwards	6.0 (5.8, 6.2)	+43%	10.3 (5.6, 14.9)	+37%	19.6 (10.7, 28.5)	+7%
Simulated PFS with visits at 8, 16 and every 4 weeks afterwards	8.0 (8.0,8.0)	+90%	9.2 (4.5-14.0)	+23%	19.9 (11.7, 28.1)	+9%
Hazard Ratio between Simulated Curves						
HR for PFS with 6, 12, 16… vs. PFS with 8, 16, 20, …	0.89 (0.83-0.95)	-11.2%	0.92 (0.87-0.97)	-8.0%	0.97 (0.95-1.00)	-2.6%

PFS: progression-free survival, HR: hazard ratio, CI: confidence interval, Diff: difference in medians.

Table 3. Summary of results on HR with and without ASM

		Without ASM						With ASM					
Parameter	**Scenario**	**HR***	**95% CI**	**% Bias**	**Cov**	**rMSE**	**SE log HR**	**HR***	**95% CI**	**% Bias**	**Cov**	**rMSE**	**SE log HR**
Base case	**True HR*=1.25**	1.15	(0.86, 1.54)	-8.2%	91.3%	0.196	0.150	1.26	(0.94, 1.67)	0.6%	96.1%	0.187	0.149
Sample size	**75**	1.15	(0.82, 1.61)	-8.4%	92.5%	0.220	0.174	1.26	(0.90, 1.75)	0.4%	95.2%	0.218	0.173
	250	1.14	(0.95, 1.37)	-8.6%	83.9%	0.148	0.095	1.25	(1.05, 1.50)	0.4%	95.5%	0.116	0.094
PFS at T_1	**0.4**	1.11	(0.83, 1.48)	-11.3%	87.2%	0.209	0.148	1.26	(0.95, 1.66)	0.6%	96.3%	0.181	0.147
	0.8	1.22	(0.90, 1.65)	-2.3%	94.9%	0.190	0.154	1.25	(0.93, 1.69)	0.3%	96.4%	0.193	0.153
Follow-up	**52 weeks**	1.14	(0.84, 1.55)	-8.8%	90.4%	0.207	0.157	1.26	(0.93, 1.70)	0.6%	95.4%	0.198	0.156
	156 weeks	1.15	(0.86, 1.54)	-7.9%	91.0%	0.192	0.148	1.26	(0.95, 1.67)	0.6%	95.8%	0.185	0.147
Progression assessment schedules + Adjusting First Visit	**Every 6 vs 8 weeks**	1.15	(0.86, 1.54)	-8.0%	91.4%	0.195	0.150	1.26	(0.95, 1.68)	0.8%	96.2%	0.188	0.150
	Every 6 vs 9 weeks	1.14	(0.85, 1.53)	-8.9%	90.5%	0.199	0.150	1.25	(0.94, 1.67)	0.1%	95.9%	0.186	0.150
Progression assessment schedules + Adjusting All Visits	**Every 6 vs every 8 weeks**							1.26	(0.95, 1.68)	0.9%	96.1%	0.188	0.149
	Every 6 vs every 9 weeks							1.26	(0.95, 1.68)	0.8%	95.7%	0.188	0.150
ATB correction scenario	**Worst-case Scenario**							1.23	(0.92, 1.63)	-1.9%	95.2%	0.182	0.150
	Linear Interpolation							1.26	(0.95, 1.68)	1.0%	96.0%	0.189	0.149
HR (comp vs index)	**0.80**	0.74	(0.55, 1.01)	-7.2%	92.4%	0.127	0.154	0.80	(0.59, 1.09)	0.1%	94.5%	0.126	0.154

		Without ASM						With ASM					
Parameter	**Scenario**	**HR***	**95% CI**	**% Bias**	**Cov**	**rMSE**	**SE log HR**	**HR***	**95% CI**	**% Bias**	**Cov**	**rMSE**	**SE log HR**
	1	0.92	(0.69, 1.24)	-7.6%	91.8%	0.157	0.151	1.00	(0.75, 1.35)	0.5%	95.4%	0.153	0.151

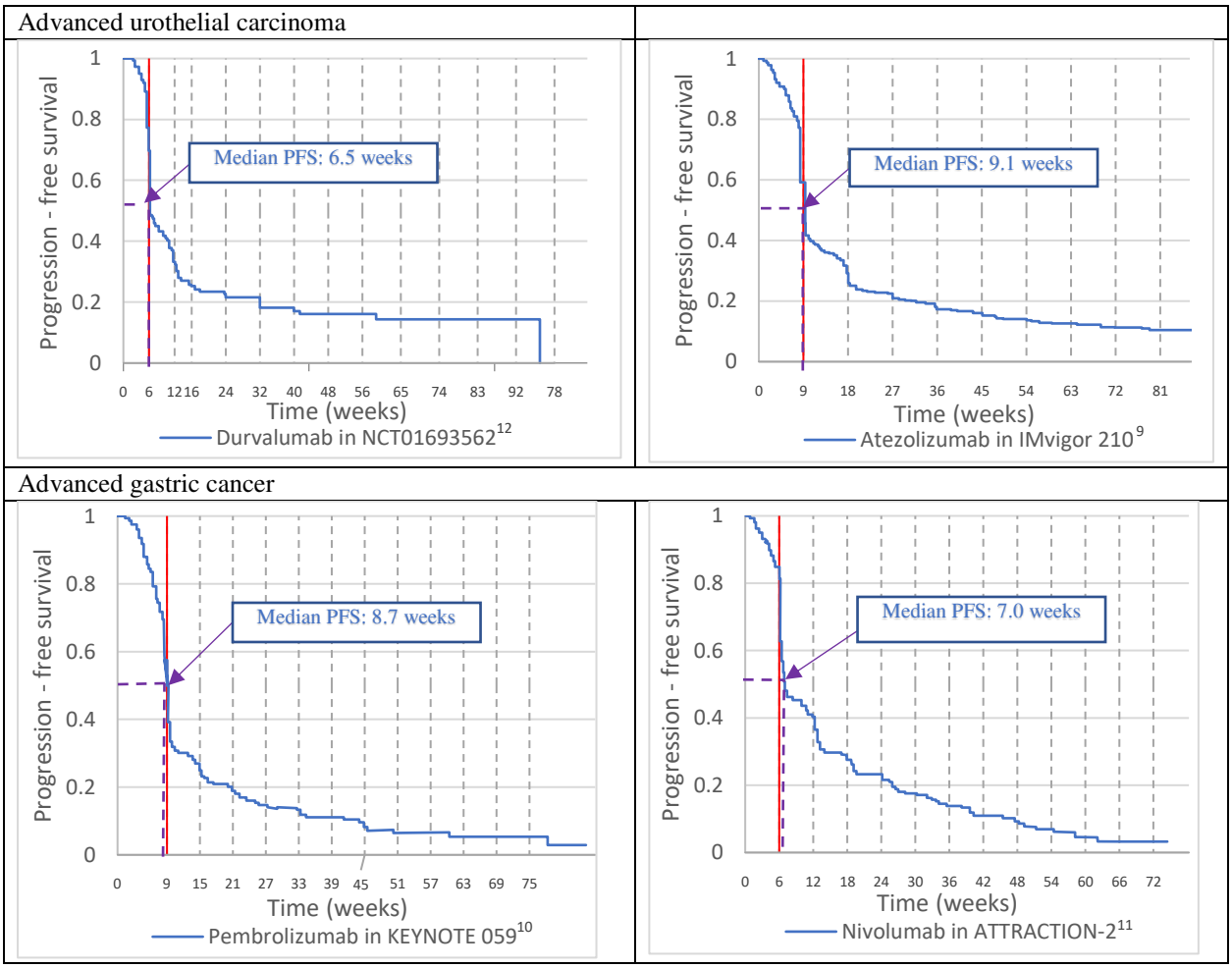
ASM: assessment-time matching, PFS, progression-free survival; HR, hazard ratio; CI, confidence interval, Cov, coverage; rMSE, root mean square error; SE, standard error; T_1: first progression assessment time in index study.

The base case scenario used a sample size of 100 patients in each study, HR=1.25, maximum follow-up 104 weeks, progression assessments at 6, 12 and every 4 weeks afterwards for index study and at 8, 16 and every 4 weeks afterwards for comparator study, and backward shift proportion calculation based on progression probability approach. Alternative scenarios were investigated by varying each parameter one at a time.

Figure Legend

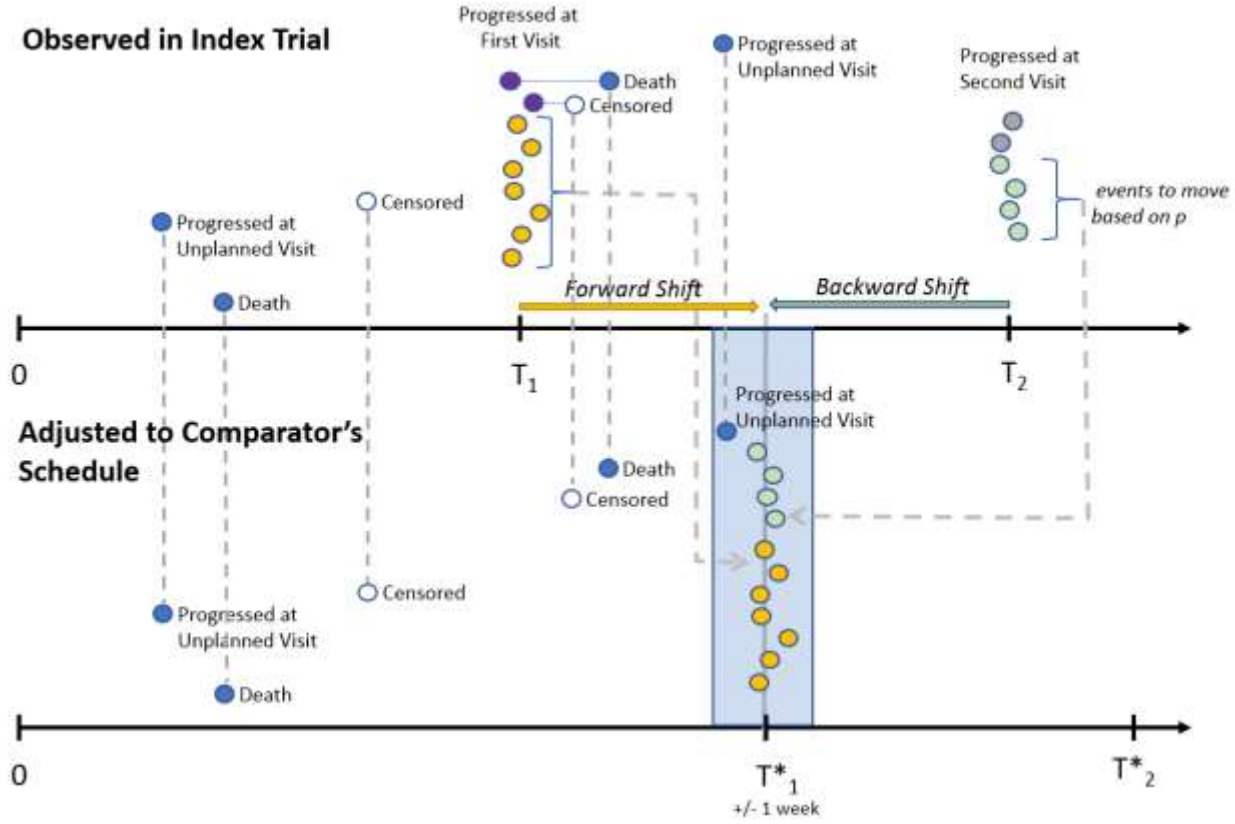
Figure 1. Progression-free survival in immunotherapy studies in aggressive carcinomas
Figure 2. Diagram Illustrating the ASM Approach
Figure 3. Description of adjustment steps at the first assessment points.
Figure 4. Empirical Progression-Free Survival Curve used as the basis for the Simulation Study.
Figure 5. Structure of the three-state model for PFS fitted to the study data and used in simulations.
Figure 6. ATB on median PFS under different progression rates at the first assessment point.
Figure 7. Empirical density of HRs across 1000 replications of the base case scenario

Figure 1. Progression-free survival in immunotherapy studies in aggressive carcinomas



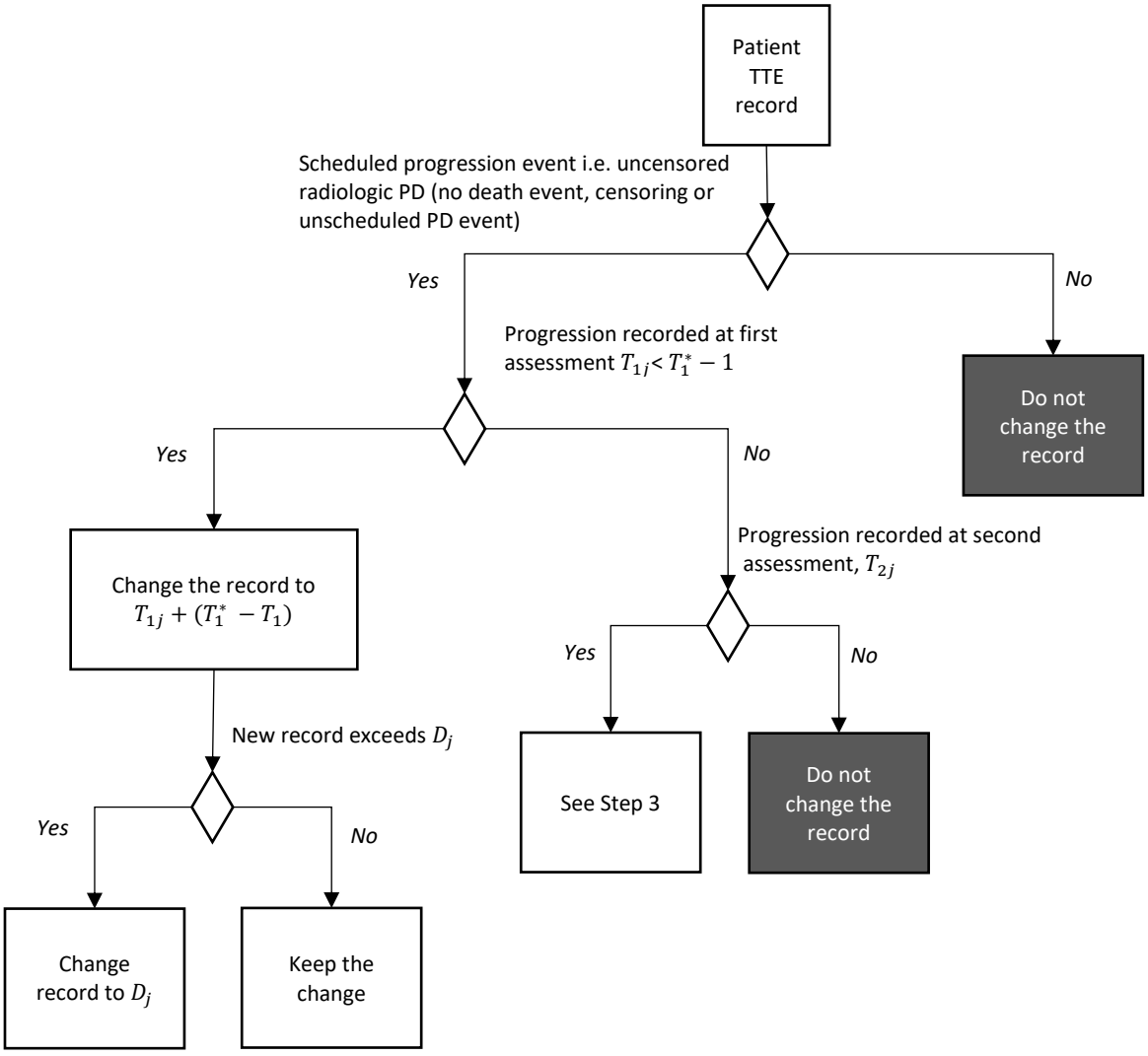
PFS: progression-free survival. Times shown reflect the schedule of progression assessment in each study. The red vertical line corresponds to the first schedule of progression assessment.

Figure 2. Diagram Illustrating the ASM Approach for matching the first progression assessment time



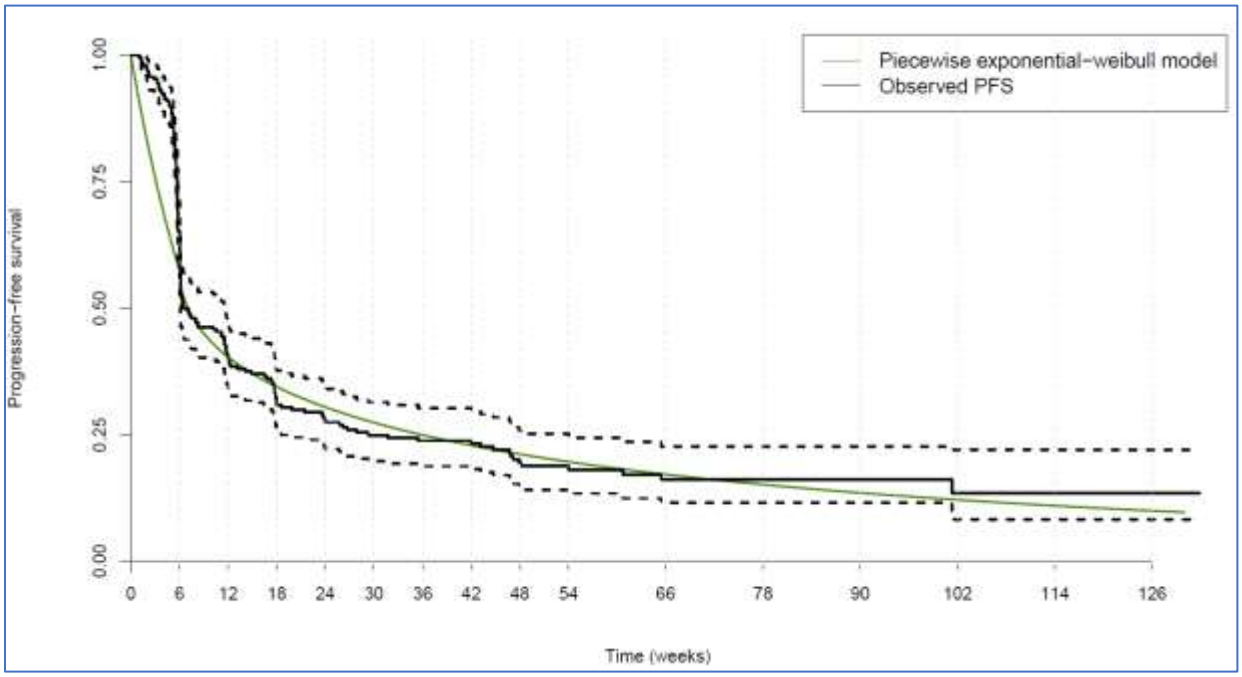
T_1: First progression assessment time in index study, T_2: Second progression assessment time in index study, T_1^*: First progression assessment time in comparator study, T_2^*: Second progression assessment time in comparator study, p: proportion of patients for whom progression could have been captured if the index study had an assessment scheduled at T_2^*. Dots have been spread out vertically for illustration purposes to avoid overlap; the vertical distance of the dots from the horizontal axis is not informative. Blue dots represent progression times recorded at unplanned visits or death times (not altered by the ASM method). White dots represent censored times (not altered by the ASM method). Yellow dots represent the progression times shifted forward by $T_1^* - T_1$ in step 1 of the ASM method. Purple dots represent the progression times that were shifted forward in step 1 of the ASM algorithm but were corrected for death or censoring in step 2 of the ASM method. Green dots represent progression times shifted backwards in step 3 of the ASM method. Grey dots represent progression times at T_2 that were not shifted backwards in step 3 of the ASM method.

Figure 3. Description of adjustment steps at the first assessment points.



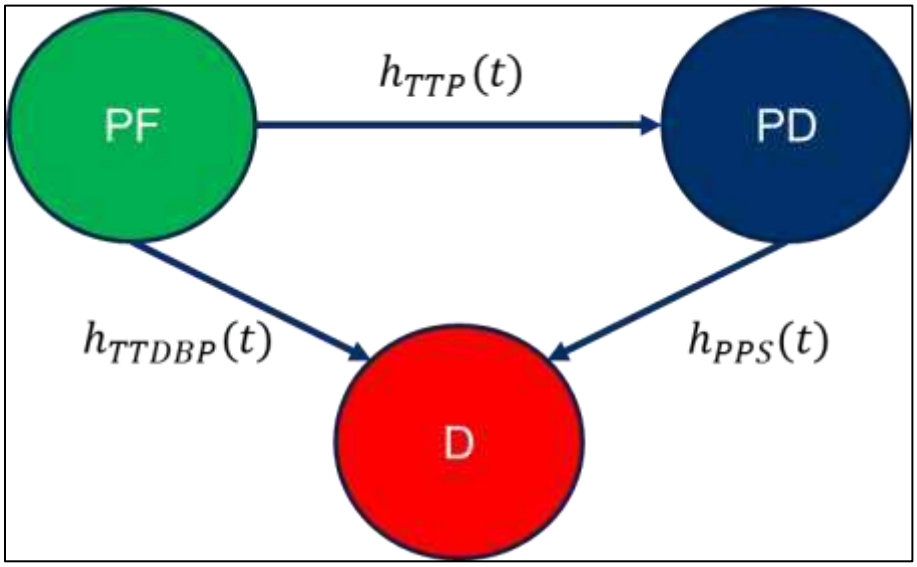
TTE: time to event, PD: progressive disease, T_{1j}: First progression assessment time for patient j in index study, T_{2j}: Second progression assessment time for patient j in index study, T_1: First progression assessment time in index study, T_1^*: First progression assessment time in comparator study, D_j: death or censoring time for patient j in the index study.

Figure 4. Empirical Progression-Free Survival Curve used as the basis for the Simulation Study.



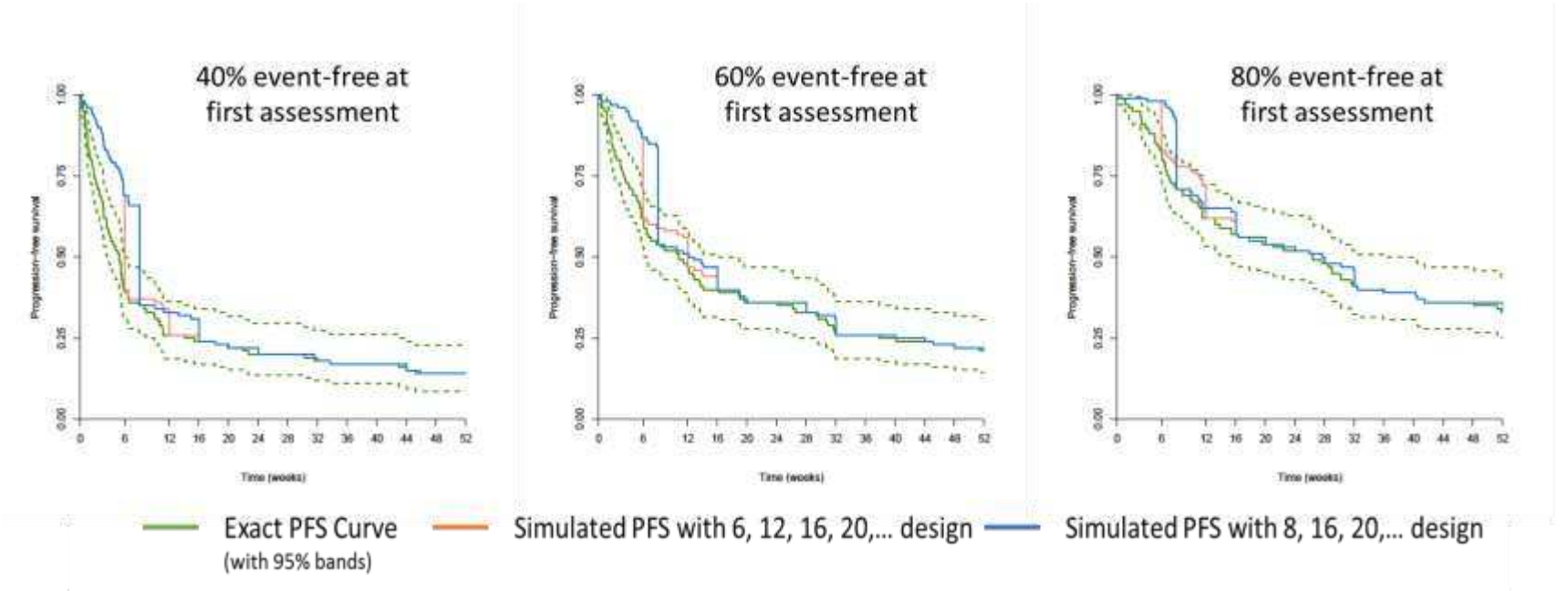
Dashed lines represent 95% confidence band around the observed PFS curve.

Figure 5. Structure of the three-state model for PFS fitted to the study data and used in simulations.



PF = Progression free; PD = Progressed disease; D = death; h(t) = hazard of transitioning from one node to the other; TTP = Time to progression; PPS = Post-progression survival; TTDBP = Time to death before progression.

Figure 6. ATB on median PFS under different progression rates at the first assessment point.



ATB: assessment-time bias, PFS: progression-free survival.

Figure 7. Empirical density of HRs across 1000 replications of the base case scenario.

HR: hazard ratio, ASM: assessment-schedule matching