

Method

Cite this article: Salmon D, Melendez-Torres GJ (2023). Clinical effectiveness reporting of novel cancer drugs in the context of non-proportional hazards: a review of nice single technology appraisals. *International Journal of Technology Assessment in Health Care*, 39(1), e16, 1–8
<https://doi.org/10.1017/S0266462323000119>

Received: 30 September 2022

Revised: 15 January 2023

Accepted: 30 January 2023

Key words:


non-proportional hazards; National Institute for Health and Care Excellence; health technology assessment; survival analysis; cancer

Author for correspondence:

*David Salmon,

E-mail: david.salmon1@nhs.net

Clinical effectiveness reporting of novel cancer drugs in the context of non-proportional hazards: a review of nice single technology appraisals

David Salmon^{1*}  and G. J. Melendez-Torres²

¹Faculty of Health and Life Sciences, University of Exeter, Devon, UK and ²Peninsula Technology Assessment Group (PenTAG), Faculty of Health and Life Sciences, University of Exeter, Devon, UK

Abstract

Objectives: The hazard ratio (HR) is a commonly used summary statistic when comparing time to event (TTE) data between trial arms, but assumes the presence of proportional hazards (PH). Non-proportional hazards (NPH) are increasingly common in NICE technology appraisals (TAs) due to an abundance of novel cancer treatments, which have differing mechanisms of action compared with traditional chemotherapies. The goal of this study is to understand how pharmaceutical companies, evidence review groups (ERGs) and appraisal committees (ACs) test for PH and report clinical effectiveness in the context of NPH.

Methods: A thematic analysis of NICE TAs concerning novel cancer treatments published between 1 January 2020 and 31 December 2021 was undertaken. Data on PH testing and clinical effectiveness reporting for overall survival (OS) and progression-free survival (PFS) were obtained from company submissions, ERG reports, and final appraisal determinations (FADs).

Results: NPH were present for OS or PFS in 28/40 appraisals, with log-cumulative hazard plots the most common testing methodology (40/40), supplemented by Schoenfeld residuals (20/40) and/or other statistical methods (6/40). In the context of NPH, the HR was ubiquitously reported by companies, inconsistently critiqued by ERGs (10/28), and commonly reported in FADs (23/28).

Conclusions: There is inconsistency in PH testing methodology used in TAs. ERGs are inconsistent in critiquing use of the HR in the context of NPH, and even when critiqued it remains a commonly reported outcome measure in FADs. Other measures of clinical effectiveness should be considered, along with guidance on clinical effectiveness reporting when NPH are present.

Introduction

Estimating clinical effectiveness of a novel treatment versus a comparator is an essential component of NICE technology appraisals (TAs). For time-to-event (TTE) outcomes such as overall survival (OS) and progression-free survival (PFS), effect size is often summarized by the hazard ratio (HR); a summary statistic describing the relative difference between two survival curves. Most commonly derived from semi-parametric Cox regression modeling, the HR depends on the assumption of proportional hazards (PH). This means that the ratio of the hazard functions of each curve should remain constant with time, or alternatively, that any changes in hazard rates over time in one curve should be accompanied by proportionate changes in the other (1). Methodologies for testing the PH assumption include visualization of log-cumulative hazard plots, which demonstrate approximately parallel lines with no crossover in the presence of PH. Alternatively, Schoenfeld residuals (2), summarized as the observed minus the expected values of the covariates at each failure time, demonstrate whether a covariate coefficient is time-dependent – this should not be the case when the PH assumption holds. Grambsch–Therneau tests are an extension of this – testing for correlation between a covariate’s Schoenfeld residual and a function of time, with a non-zero correlation suggesting PH violation (3).

The PH assumption in oncology trials

Novel oncology drugs include targeted and immuno-oncology (IO) therapies. Targeted therapies selectively inhibit the growth of cancer cells by interfering with enzymes or cell signaling pathways (4). Examples of targets include DNA-repair enzymes (e.g., poly-ADP ribose polymerase inhibitors olaparib and niraparib), proteins involved in cell division (e.g., cyclin-dependent kinase 4 and 6 inhibitors palbociclib and abemaciclib), and cytoplasmic tyrosine kinase domains of various receptors (e.g., anaplastic lymphoma kinase, fms-like tyrosine kinase

3, epidermal growth factor receptor, and brigatinib, gilteritinib and osimertinib, respectively). IO therapies such as immune checkpoint inhibitors function by stimulating the immune system to destroy cancerous cells (4). Examples of targets include inhibitors of the down-regulators of immune T-cell activity such as programmed death receptor 1 (pembrolizumab and nivolumab), and its ligand (atezolizumab, avelumab), and cytotoxic T-lymphocyte-associated antigen 4 (ipilimumab).

The mechanisms of action of these novel drugs vary both within class, and with the traditional chemotherapies to which they are often compared in randomized controlled trials (RCTs). As a result, the shapes of survival curves in TAs often vary considerably between intervention and comparator, leading to violation of the PH assumption - or non-proportional hazards (NPH). Patterns seen include the following (5): (i) Delayed treatment effects, often seen with IO therapies as their impact on stimulating the immune system takes time to build; (ii) Treatment waning effects, often a result of cancer cells developing mutations or molecular bypass pathways which with time allow them to “escape” direct treatment effects or evade the IO-boosted immune system; (iii) Durable survival or cure, whereby survival benefit is maintained in the longer term, even after treatment cessation; and (iv) Crossing hazards, whereby survival curves of intervention and comparator cross (Figure 1).

There are also artifactual reasons (independent of true treatment effects) that may cause an apparent violation of the PH assumption. For example, treatment switching after disease progression can confound OS by diluting observed treatment benefits, impacting on hazard patterns. Alternatively, “pseudo-progression” (an initial perceived increase in tumor volume due to infiltration with immune cells) can occur on commencement of IO therapy (6). This is transient, but if not recognized can be mistaken for true disease progression and impact PFS curves and hazard patterns. Similarly impacting PFS, an apparent (but false) delayed treatment effect can be observed as an artifact of measurement schedule. For example, an initial non-divergence of curves may be seen due to the first follow up not being for some time after randomization. Finally, the presence of subgroups of

patients with differing response to treatment can have considerable impact. In the IPASS (7) trial (gefitinib vs. carboplatin-paclitaxel for pulmonary adenocarcinoma), crossing hazards were seen in the raw analysis, but resolved when patients were stratified by EGFR status. This treatment effect heterogeneity may be due to unobserved effect modifiers.

The problem with the hazard ratio

In the context of PH, the HR is a meaningful summary statistic describing differences in treatment effect between two groups. However, it is a relative measure with no absolute component, thus requiring other statistics (e.g., percentile or landmark survival) to furnish it with clinical meaning. Some have argued it is unintuitive and poorly understood by clinicians (8), while others contest its value from a causal inference perspective. As described by Stensrud et al. (9), in RCTs individuals at high and low risk of experiencing an event should (theoretically) be evenly distributed between treatment and comparator groups at baseline. When a treatment is effective this balance is lost with time, as not only will more individuals survive for longer in the treatment group, but it will accrue a greater proportion of high-risk individuals, who would have died sooner in the comparator group. The result is selection bias; whereby the HR summarizes not only the difference in treatment effect, but also the growing differences in characteristics (known and unknown) between the two populations.

When NPH are observed, whether this is due to true treatment effect or artifact, the HR (by definition) varies with time. Therefore, reporting an “overall” HR loses meaning as it does not describe how events are distributed through the trial follow-up period, giving no information regarding delayed effects, waning effects, or crossing hazards. Moreover, its value will change depending on the (somewhat arbitrary) duration of the trial. Methods within the framework of Cox regression to accommodate NPH have been used including time-dependent covariates (10), or the use of multiple piecewise HRs (11). However, there are practical difficulties with implementing the former (1) (which are seldom used in trial reporting), while the latter is subject to the biases described above.

Alternatives to the hazard ratio

In addition to the HR, many trials report cross-sectional measures of survival including percentile (i.e., the timepoint when x percent of patients have experienced the event, fifty percent being the median) and landmark (i.e., the survival probability at a given time point) survival. Uno et al (12) discuss using ratios or differences in these to describe treatment effect, based on prespecified, “clinically meaningful” milestones. However, these are snapshots of a single point in time, and can be misleading when quoted in isolation in the context of NPH. For example, in TA620 (13) (Figure 2), median survival for intervention and comparator is approximately equal (thirty months); but while this gives the impression of equal treatment effect, the curves clearly diverge beyond this. Similarly, thirty-month landmark survival is equal between groups, but this is not the case later. Which milestone is most informative for decision making?

Alternatively, the log-rank test is a commonly reported non-parametric hypothesis test of treatment effect difference between arms (14). This uses a test statistic derived from the ranks of survival times between two populations, compared with a chi-square distribution. The resulting p-value determines whether evidence exists to reject the null hypothesis of no treatment effect

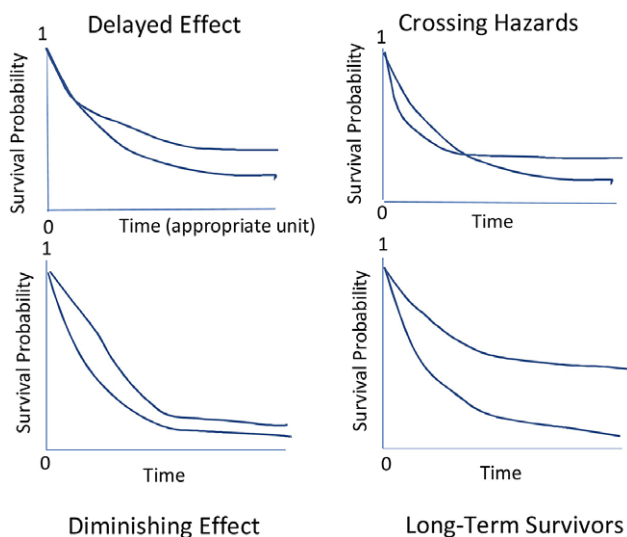


Figure 1. Examples of survival curves demonstrating non-proportional hazards. Clockwise from top left: Delayed treatment effect, crossing hazards, long term survival, and diminishing (treatment waning) effect. Reproduced with permission from Ananthakrishnan et al. (5).

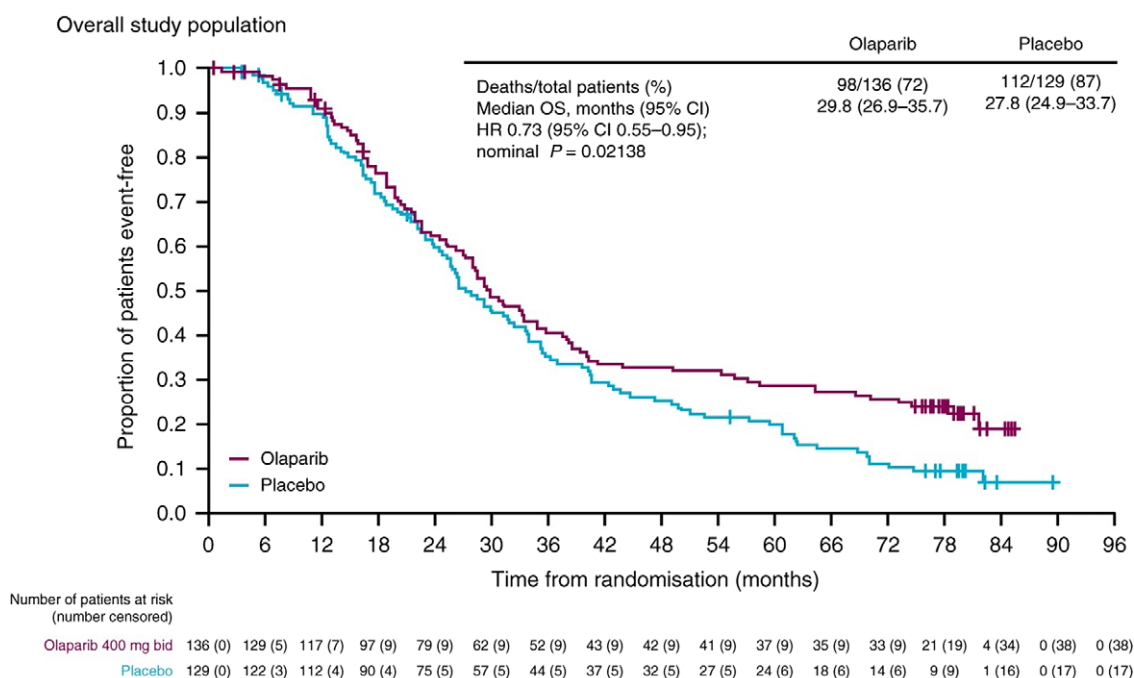


Figure 2. An example of a delayed treatment effect.

Note how medians are similar, before the curves diverge from around 30 months onwards. Taken from TA620 (13), original image from the Study-19 trial (40) (open access article).

to a predefined significance level. While log-rank is statistically valid under NPH, it loses power in this context (15), increasing the probability of type two error. In response, many authors have advocated for weighted log-rank tests, whereby different parts of the survival curves are given different emphasis. For example, early-emphasis tests (e.g., Wilcoxon (16)) may be more useful in the context of treatment waning effects, and late-emphasis tests in the context of delayed treatment effects. Others have suggested various combination tests (15). Unlike the HR, it does not describe magnitude of treatment effect; a highly significant p-value could represent strong evidence for a small difference in survival.

Restricted mean survival time (RMST) is an alternative measure, which is neatly summarized as the mean survival up to a given time point (t) and represents the area under the Kaplan–Meier curve up until t (17). As a summary of treatment effect magnitude, the ratio or difference in RMST can be reported between two curves. RMST is not dependent on the PH assumption, and is considered by many to be more intuitive than hazard-based measures (18). Unlike the median, it summarizes the entire curve up to time t and can therefore describe changes after fifty percent survival probability is crossed. However, since RMST gives equal weight to the later part of the curve where there are fewer subjects at risk, it can be more uncertain - although this can be reflected in broader confidence intervals (1). Like milestone estimates, there is subjectivity in defining t , which should therefore be prespecified to avoid selection bias (12). In comparison to the HR, RMST has been shown to be a more conservative measure of clinical effectiveness in oncology trials (19).

Relevance to NICE technology appraisals

With the recent expansion of novel oncology therapies, violation of the PH assumption is being observed with increasing frequency (1). This has implications for various aspects of TAs, including clinical

effectiveness reporting, indirect treatment comparison (ITC) methodology, and survival extrapolation for economic modeling. While technical support document fourteen (TSD14) (20) recommends PH testing as routine in NICE TAs with a view to the latter two aspects, there is limited formal guidance on best reporting of clinical effectiveness in the context of NPH.

The goal of this study is therefore to understand how pharmaceutical companies, evidence review groups (ERGs) and appraisal committees (ACs) evaluate and report clinical effectiveness in NICE TAs in the presence of NPH. By reviewing TAs of novel cancer therapies over a two-year period, the aim is to understand: to what extent PH testing is performed, the testing methods used, to what degree the presence of NPH influences clinical effectiveness reporting, and how ERGs and ACs discuss and respond to these issues.

Methods

The methodology used for this review is similar to that used for other reviews of NICE TAs (21). A complete list of appraisals was obtained from the NICE website (22). Single technology appraisals (STAs), including Cancer Drugs Fund (CDF) reviews, of targeted and IO cancer treatments published between 1 January 2020 and 31 December 2021 were included. Appraisals of chemotherapies and hormonal treatments were considered outside the scope as they are less likely to contain NPH. For CDF reviews, priority was given to the review itself, but if full information could not be obtained then the original appraisal was considered in addition. In this instance both the original appraisal and the review were considered as one appraisal.

For simplicity, the focus of data extraction for each TA was limited to OS and PFS in the pivotal RCT presented by the company. Therefore, TAs which did not contain an RCT, for example, those presenting single-arm studies or relying solely on ITC, were

excluded. Multiple technology appraisals were excluded on this basis, as the majority rely on meta-analysis. OS and PFS were chosen as these outcomes are most commonly used to evaluate clinical effectiveness, as well as being used for extrapolation and economic modeling. If PFS was not reported as an outcome, a closely related measure such as disease-free survival (DFS) was considered in its place.

Company submissions (CS), ERG reports, and final appraisal determinations (FAD) for each eligible appraisal were downloaded from the web pages in [Table 1](#). This study used thematic analysis to search across these documents in order to inductively identify, analyze and report repeating patterns. The “six steps” defined by Braun and Clarke (23) were followed. Step one was to familiarize oneself with the data by reading through (and making brief notes on) the aforementioned documents for each TA. Second, initial codes were generated. These were collated into a data extraction form comprising mostly binary or multiple-choice responses (Supplementary Material – [Table 2](#)) and tabulated on an Excel spreadsheet. Third, an initial set of themes were developed, which were then reviewed and refined for accuracy (step four) following a second review of the documents for each TA. Additionally at this stage, a significant amount of free text was captured to explore discussions amongst the company, ERG and committees on issues pertaining to NPH. Step five involved defining the final themes, while step six was the production of the narrative and manuscript.

Results

A total of seventy-one STAs assessing cancer immunotherapies or targeted treatments were identified in 2020 and 2021 from the NICE website. Thirty-one STAs were excluded: Of these eighteen were terminated appraisals and ten did not include a comparative pivotal trial. The remaining three exclusions were a rapid review, a rediscussion of an old appraisal due to a change in treatment pathway, and an update of a 2018 appraisal. The remaining forty appraisals, eleven of which were CDF reviews, were considered and are listed in [Table 1](#) (full references in the Supplementary Material – [Table 1](#)). These included treatments for hematological, pulmonary, breast, renal/urothelial, esophageal, ovarian, head and neck, colorectal, hepatocellular and dermatological malignancy. Key themes are described and explored using the sub-headings below.

Issues pertaining to PH testing: Frequency, methodology, discussions

PH testing was carried out in 39/40 company submissions where an HR was used as an outcome measure. Of these, it was ubiquitously reported in cost-effectiveness sections to inform survival extrapolation methodology. However, only 10/40 submissions reported this in the clinical effectiveness section; and in the majority of these it was done to inform ITC methodology rather than to support or dispute the validity of the HR as an outcome measure.

After engagement with the ERG, log-cumulative hazard plots were the most frequently used tool (40/40) for testing the PH assumption. In some cases, this was supplemented by Schoenfeld residual plots (20/40), and Grambsch–Therneau tests (4/40). On two occasions the ERG requested further testing with H-H plots during clarification (24;25). In 16/40 TAs, visual inspection of log-cumulative hazard plots alone was felt to be sufficient.

In 3/40 cases (26–28) the ERG and company disagreed on the results of PH testing. In two of these, the ERG critiqued the

company’s use of log-cumulative hazard plots alone for decision making (26;28). For example, in TA619 (26), the ERG commented that decision making through visual inspection of log-cumulative hazard plots alone was subjective and requested the company perform additional further testing (e.g., using Schoenfeld residuals). This was at odds with several other TAs (e.g., TA668 (29)) where log-cumulative hazard plots alone were felt to be sufficient. Perhaps conversely, in TA736 (30) the company noted that for OS Schoenfeld residual testing did not provide enough evidence to reject the PH assumption, but still deemed it violated based on visual inspection of log-cumulative hazard plots, and the differing mechanisms of action between intervention and comparator. There was therefore some inconsistency with what was deemed necessary to test the PH assumption.

In several cases, the ERG critiqued the company’s use of log(time) on the *x*-axis of log-cumulative hazard plots (27;31;32), requesting time as an alternative. For example, in TA629 (27), the company presented plots using log(time). Despite the lines crossing they were deemed otherwise parallel and therefore PH was assumed. The ERG stated: “an assessment of proportional hazards should be of the log-cumulative hazard functions against time, and a plot against log(time) was rightly criticized because the long-term difference is compressed on the log(time) scale” (committee papers, ERG report p24). Based on this plot, the ERG rejected the PH assumption. In reports where plots were available to review, the vast majority (27/31) presented log(time) without criticism.

Reporting and criticism of the HR in the context of NPH

In 28/40 cases, the PH assumption was deemed to be violated in the pivotal trial for key outcomes of OS and/or PFS, either in the initial CS, or following critique by the ERG. In all cases, the HR was still reported as an outcome measure in the CS, along with other measures including log-rank testing, median TTE, and other cross-sectional measures such as percentile or landmark survival. While ERGs performed thorough evaluations of the PH assumption when critiquing ITC and survival extrapolation methodology, criticism of the use of the HR as a measure of clinical effectiveness in the presence of NPH was less consistent (10/28). In these cases where the use of HR was critiqued, it tended to be a straightforward acknowledgement of the limitations, rather than any deeper analysis of the alternatives (Supplementary Material – [Box 1](#)). Notably, whether critiqued in the committee papers or not, the HR continued to be widely quoted in FADs as a measure of clinical effectiveness (23/28) without any mention of PH assumption violation.

Use of alternative measures to the HR

In the context of NPH, measures of clinical efficacy ubiquitously reported in addition to the HR included log-rank tests, median and other percentile TTE (where estimable), and landmark survival. RMST was the only other measure used, but was only explored in three TAs (13;33;34). In TA620 (13) a delayed treatment effect was noted ([Figure 2](#)) for OS. Median survival was a poor estimate as 50 percent of patients had died prior to separation of the curves, and the ERG suggested: “restricted means analysis gives a more informative and reliable estimate of survival benefit compared with the HR”, and that visualization of survival curves may give the best estimate of treatment effects, followed by event rates at certain timepoints. However, although the HR was presented “for completion”, this was still the key measure reported in the FAD,

Table 1. Technology appraisals included in the review

Title	Year	TA	Type	ERG	Pivotal trial
Olaparib for maintenance treatment of relapsed platinum-sensitive ovarian, fallopian tube or peritoneal cancer	2020	TA620	STA	BMJ-TAG	Study19, SOLO-2
Palbociclib with fulvestrant for treating hormone receptor-positive, HER2-negative, advanced breast cancer	2020	TA619	STA	LRiG	PALOMA3
Lenalidomide with rituximab for previously treated follicular lymphoma	2020	TA627	STA	KSR	AUGMENT
Obinutuzumab with bendamustine for treating follicular lymphoma after rituximab	2020	TA629 (TA472)	CDF	SCHARR-TAG	GADOLIN
Trastuzumab emtansine for adjuvant treatment of HER2-positive early breast cancer	2020	TA632	STA	KSR	KATHERINE
Atezolizumab with nab-paclitaxel for untreated PD-L1-positive, locally advanced or metastatic, triple-negative breast cancer	2020	TA639	STA	LRiG	IMPASSION130
Atezolizumab with carboplatin and etoposide for untreated extensive-stage small-cell lung cancer	2020	TA638	STA	KSR	IMPOWER133
Gilteritinib for treating relapsed or refractory acute myeloid leukemia	2020	TA642	STA	SCHARR-TAG	ADMIRAL
Brentuximab vedotin in combination for untreated systemic anaplastic large cell lymphoma	2020	TA641	STA	KSR	ECHELON2
Avelumab with axitinib for untreated advanced renal cell carcinoma	2020	TA645	STA	LRiG	JAVELIN RENAL 101
Polatuzumab vedotin with rituximab and bendamustine for treating relapsed or refractory diffuse large B-cell lymphoma	2020	TA649	STA	KSR	GO29365
Pembrolizumab with axitinib for untreated advanced renal cell carcinoma	2020	TA650	STA	SHTAC	KEYNOTE426
Osimertinib for treating EGFR T790M mutation-positive advanced non-small-cell lung cancer	2020	TA653 (TA416)	CDF	LRiG	AURA3
Nivolumab for advanced squamous non-small-cell lung cancer after chemotherapy	2020	TA655 (TA483)	CDF	LRiG	CHECKMATE 017
Isatuximab with pomalidomide and dexamethasone for treating relapsed and refractory multiple myeloma	2020	TA658	STA	SCHARR-TAG	ICARIA MM
Pembrolizumab for untreated metastatic or unresectable recurrent head and neck squamous cell carcinoma	2020	TA661	STA	LRiG	KEYNOTE048
Venetoclax with obinutuzumab for untreated chronic lymphocytic leukemia	2020	TA663	STA	Warwick Evidence	CLL14
Atezolizumab with bevacizumab for treating advanced or unresectable hepatocellular carcinoma	2020	TA666	STA	SCHARR-TAG	IMBRAVE
Encorafenib plus cetuximab for previously treated BRAF V600E mutation-positive metastatic colorectal cancer	2021	TA668	STA	Warwick Evidence	BEACON-CRC
Brigatinib for ALK-positive advanced non-small-cell lung cancer that has not been previously treated with an ALK inhibitor	2021	TA670	STA	LRiG	ALTA-1 L
Niraparib for maintenance treatment of advanced ovarian, fallopian tube and peritoneal cancer after response to first-line platinum-based chemotherapy	2021	TA673	STA	BMJ-TAG	PRIMA
Lenalidomide maintenance treatment after an autologous stem cell transplant for newly diagnosed multiple myeloma	2021	TA680	STA	PENTAG	MYELOMA XI
Pembrolizumab with pemetrexed and platinum chemotherapy for untreated, metastatic, non-squamous non-small-cell lung cancer	2021	TA683 (TA557)	CDF	PENTAG	KEYNOTE189
Nivolumab for adjuvant treatment of completely resected melanoma with lymph node involvement or metastatic disease	2021	TA684 (TA558)	CDF	BMJ-TAG	CHECKMATE 238
Ribociclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy	2021	TA687 (TA593)	CDF	BMJ-TAG	MONALEESA 3
Acalabrutinib for treating chronic lymphocytic leukemia	2021	TA689	STA	SCHARR-TAG	ELEVATE-TN
Carfilzomib with dexamethasone and lenalidomide for previously treated multiple myeloma	2021	TA695	STA	BMJ-TAG	ASPIRE
Olaparib plus bevacizumab for maintenance treatment of advanced ovarian, fallopian tube or primary peritoneal cancer	2021	TA693	STA	BMJ-TAG	PAOLA-1
Pembrolizumab for treating locally advanced or metastatic urothelial carcinoma after platinum-containing chemotherapy	2021	TA692 (TA519)	CDF	Warwick Evidence	Keynote 045
Atezolizumab monotherapy for untreated advanced non-small-cell lung cancer	2021	TA705	STA	HERU/HSRU	IMPOWER-110

(Continued)

Table 1. (Continued)

Title	Year	TA	Type	ERG	Pivotal trial
Nivolumab for previously treated unresectable advanced or recurrent esophageal cancer	2021	TA707	STA	PENTAG	ATTRACTION-3
Pembrolizumab for untreated metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency	2021	TA709	STA	BMJ-TAG	KEYNOTE-177
Nivolumab for advanced non-squamous non-small-cell lung cancer after chemotherapy	2021	TA713 (TA484)	CDF	LRIG	CHECKMATE 057
Nivolumab with ipilimumab and chemotherapy for untreated metastatic non-small-cell lung cancer	2021	TA724	STA	CRD/CHE	CHECKMATE 9LA
Abemaciclib with fulvestrant for treating hormone receptor-positive, HER2-negative advanced breast cancer after endocrine therapy	2021	TA725 (TA579)	CDF	BMJ-TAG	Monarch 2
Pembrolizumab with platinum- and fluoropyrimidine-based chemotherapy for untreated advanced esophageal and gastro-esophageal junction cancer	2021	TA737	STA	PENTAG	KEYNOTE-590
Nivolumab for treating recurrent or metastatic squamous cell carcinoma of the head and neck after platinum-based chemotherapy	2021	TA736 (TA490)	CDF	KSR	CHECKMATE 141
Atezolizumab for untreated PD-L1-positive advanced urothelial cancer when cisplatin is unsuitable	2021	TA739 (TA492)	CDF	SHTAC	IMvigor130
Nivolumab for adjuvant treatment of resected esophageal or gastro-esophageal junction cancer	2021	TA746	STA	SchARR-TAG	CHECKMATE 577
Mogamulizumab for previously treated mycosis fungoides and Sézary syndrome	2021	TA754	STA	KSR	MAVORIC

Note for CDF reviews the original appraisal is included in parenthesis. A fully referenced copy of this table can be found in the [supplementary material](#).

Abbreviation: TA = Technology appraisal; ERG = Evidence Review Group; PH=Proportional Hazard; STA = Single Technology Appraisal; CDF = Cancer Drugs Fund review; PENTAG = Peninsula Technology Assessment Group (University of Exeter); SchARR-TAG = School of Health and Related Research Technology Assessment Group (University of Sheffield); BMJ-TAG = British Medical Journal Technology Assessment Group; LRIG = Liverpool Reviews and Implementation Group (University of Liverpool); SHTAC=Southampton Health Technology Assessment Centre (University of Southampton); KSR = Kleijnen Systematic Reviews; HERU/HSRU=Health Economics Research Unit and Health Services Research Unit (University of Aberdeen); CRD/CHE = Centre for Reviews and Dissemination (CRD) and Centre for Health Economics (CHE) (University of York).

with no mention of NPH. Similarly, in TA638 (34), a restricted mean analysis was used to assess conformity to end of life criteria, but again was not reported as a primary measure of clinical effectiveness in the FAD. In the CDF review TA484/TA713 (33) the company reported RMST both for the pivotal trial and the indirect treatment comparison, with the ERG noting this being the first use in the ITC setting. Interestingly, the original appraisal (TA484) was one of the few TAs where an HR was available but not mentioned in the FAD, with landmark and median survival quoted to support clinical efficacy claims. However, in the CDF review (TA713) FAD, a statistically significant (albeit confidential) HR was reported, despite no obvious changes to PH testing outcomes.

Discussion

These results confirm that violation of the PH assumption is seen in a majority of recent NICE appraisals of targeted and IO cancer treatments. While PH testing is commonplace, the results demonstrate inconsistency in how companies and ERGs assess the PH assumption, with some preferring visual inspection of log-cumulative hazard plots and others preferring formal statistical testing. There is also some variability as to whether log-cumulative hazard plots should be plotted against (log-time) or (time). Given that this method of assessing the PH assumption depends on visual inspection (which is inherently subjective), it is important that this choice is consistent amongst appraisals.

In most cases, PH testing was done to inform ITC and survival extrapolations rather than to inform clinical effectiveness reporting. As a result, use of the HR for reporting clinical effectiveness was inconsistently critiqued by the companies and ERG. When company submissions or ERG reports are read in order, it can

sometimes appear that clinical effectiveness conclusions are drawn before PH testing has been performed, as PH testing is often only discussed in the subsequent cost-effectiveness section. As a result, in the presence of NPH the HR was still ubiquitously reported as a measure of clinical effectiveness, and was reported in the majority of FADs, even when its use was critiqued by the ERG. There is a sense that the HR continues to be reported by convention, rather than as a meaningful parameter.

Does this matter? Firstly, there are some who argue that even when the PH assumption is violated, the HR is still a useful measure of “overall” treatment effect, or as a weighted average of the true hazard ratios over an entire follow up period (35). However, this is controversial, with many authors highlighting aforementioned issues with confounding (9;36). More importantly, taking the prior example of TA620 (13) (Figure 2), which demonstrates a significant HR of 0.73 (95% confidence interval 0.55–0.95) in favor of treatment but similar median survivals as an example: How can such an “overall” HR be meaningfully interpreted when 50 percent of patients do not get any benefit? Therefore, although some have stated that, to the best of their knowledge, “the use of HRs...as primary analysis tools has not impeded the development, testing, and acceptance of effective oncologic therapies” (37), it is clear that in the context of NPH, the HR is: (i) Lacking meaning as a measure of the magnitude of treatment effect, and (ii) Prone to bias. Moreover, can reporting of the HR in this instance actually be misleading? Previous vignette studies have demonstrated that a trial’s choice of measure to describe clinical effectiveness can bias clinicians’ willingness to prescribe, and how information on treatment choices is presented to patients (8;38). It is plausible this could influence appraisal committees too, and it could therefore be hypothesized that the reporting of an HR in the context of NPH

not only provides no useful additional information beyond that offered by non-parametric (and statistically valid) measures such as log-rank and percentile/landmark-based measures, but could, in fact, have an adverse impact on decision making.

Secondly, the choice of methodology used both in indirect treatment comparison and modeling and extrapolating survival curves for the economic analysis is determined by the presence or absence of NPH (39). Therefore, it could be argued clinical effectiveness reporting based on trial data alone is somewhat academic from an HTA perspective, as independently fitted parametric models can be used for mean survival estimates used in economic modeling (39). However, despite an increasing reliance on extrapolation and clinical expert opinion (particularly for the immature data submitted in many TAs (21)), surely we still need *some* trial-based evidence for clinical effectiveness. The question then is how this data can be reported in a way that is fair and consistent.

One commonly reported barrier to deeper exploration of alternative methods of clinical effectiveness reporting is the requirement to prespecify the primary analysis. For example, in TA619 (26), the company defended their use of a Cox PH model on these grounds, declining the ERG's request to provide alternative estimators. The uncertainty as to the presence or absence of PH before data collection has implications for choice of statistical testing planned, power calculations, timing of interim and futility analyses, and communicating the results with clinicians and the general public (15). Another barrier is perhaps the lack of clear guidance from HTA agencies on appropriate alternatives.

To our knowledge, this is the first study addressing PH testing and clinical effectiveness reporting practices in NICE TAs. Limitations include the reliance on written summaries of meetings, which may not accurately reflect the actual conversations that took place. While in some committee papers numerical data such as survival curves and summary statistics were redacted the key information regarding testing and identification of NPH, and discussions in this context remained obtainable from the text.

To conclude: Although not ubiquitous, several HTA agencies internationally (of which NICE is one) provide guidelines and recommendations on PH testing (1). However, in the UK, there is a lack of consistency amongst companies and ERGs both in how the PH assumption is tested (with some valuing visual inspection over formal statistical testing, or vice versa), and how the HR is critiqued in the context of NPH. Moreover, any critique does not necessarily result in a change to reporting habits; the seemingly routine reporting of the HR in committee papers and FADs should be reconsidered.

The key issue, therefore, is how NPH are managed in terms of clinical effectiveness reporting, and the value of providing the HR or alternative measures in this context. When reporting magnitude of treatment effect, some TAs recommended quoting sequential percentile or landmark estimates, with or without RMST. However, RMST was only used in a minority of appraisals and, despite some arguing it should be more widely reported in NICE TAs (1), has its own aforementioned limitations. Indeed, all single summary statistics have limitations, but perhaps the log-rank test is the most informative and least misleading in this situation; it is valid under NPH, and can tell us if there is reliable evidence of a difference between the entirety of the two arms. To ensure fairness of process, the production of guidance or standards on clinical effectiveness reporting in the context of NPH should be considered by NICE.

Supplementary materials. To view supplementary material for this article, please visit <http://doi.org/10.1017/S0266462323000119>.

Funding statement. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest statement. Professor G.J Melendez-Torres is the chief investigator of an NIHR grant to provide HTA advice to NICE.

References

1. **Monnickendam G, Zhu M, McKendrick J, Su Y.** Measuring survival benefit in health technology assessment in the presence of nonproportional hazards. *Value Heal [Internet]*. 2019;22(4):431–438. doi: 10.1016/j.jval.2019.01.005.
2. **Schoenfeld D.** Partial residuals for the proportional hazards regression model. *Biometrika*. 1982;69(1):239–241.
3. **Metzger SK.** Proportionally less difficult?: reevaluating keele's "proportionally difficult". *Polit Anal*. 2022;31:156–163.
4. **Seebacher NA, Stacy AE, Porter GM, Merlot AM.** Clinical development of targeted and immune based anti-cancer therapies. *J Exp Clin Cancer Res*. 2019;38:156.
5. **Ananthakrishnan R, Green S, Previtali A, et al.** Critical review of oncology clinical trial design under non-proportional hazards. *Crit Rev Oncol Hematol [Internet]*. 2021;162:103350. doi: 10.1016/j.critrevonc.2021.103350.
6. **Ma Y, Wang Q, Dong Q, Zhan L, Zhang J.** How to differentiate pseudo-progression from true progression in cancer patients treated with immunotherapy. *Am J Cancer Res [Internet]*. 2019;9(8):1546–1553. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/31497342%0A>; <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6726978>.
7. **Mok TS, Wu Y, Thongprasert S, et al.** Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *N Engl J Med [Internet]*. 2009;361(10):947–957. doi: 10.1056/NEJMoa0810699.
8. **Saad ED, Zalberg JR, Pcron J, et al.** Understanding and communicating measures of treatment effect on survival: Can we do better? *J Natl Cancer Inst*. 2018;110(3):232–240.
9. **Stensrud MJ, Aalen JM, Aalen OO, Valberg M.** Limitations of hazard ratios in clinical trials. *Eur Heart J*. 2019;40(17):1378–1383.
10. **Fisher LD, Lin DY.** Time-dependent covariates in the Cox proportional-hazards regression model. *Annu Rev Public Health*. 1999;20:145–157.
11. **Roychoudhury S, Anderson KM, Ye J, Mukhopadhyay P.** Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group. *Stat Biopharm Res*. 2021;13:1–15.
12. **Uno H, Claggett B, Tian L, et al.** Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32(22):2380–2385.
13. **National Institute for Health and Care Excellence (NICE).** Olaparib for maintenance treatment of relapsed platinum-sensitive ovarian, fallopian tube or peritoneal cancer. Technology appraisal guidance [TA620] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta620>.
14. **Collett D.** *Modelling survival data in medical research*. 3rd ed. London: Chapman and Hall/CRC; 2015.
15. **Duke University, US Food and Drug Administration.** Public workshop: Oncology clinical trials in the presence of non-proportional hazards. 2018. Available from: <https://www.youtube.com/watch?v=npufYAHeoxk&t=3288s>.
16. **Gehan EA.** A generalized two-sample wilcoxon test for doubly censored data. *Biometrika [Internet]*. 1965;52(3/4):650–653. Available from: <http://www.jstor.org/stable/2333721>.
17. **Royston P, Parmar MKB.** Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13:152.
18. **Wei Y, Royston P, Tierney JF, Parmar MKB.** Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: Application to individual participant data. *Stat Med*. 2015;34(21):2881–2898.

19. **Liang F, Zhang S, Wang Q, Li W.** Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. *Ann Oncol.* 2018;**29**(5):1320–1324.
20. **Latimer N.** NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. Decis Support Unit [Internet]. 2011. Available from: <http://www.nicedsu.org.uk/NICEDSUTSDSurvivalanalysis.updatedMarch2013.v2.pdf>.
21. **Bell Gorrod H, Kearns B, Stevens J, et al.** A review of survival analysis methods used in NICE technology appraisals of cancer treatments: Consistency, limitations, and areas for improvement. *Med Decis Mak.* 2019;**39**(8):899–909.
22. **National Institute for Health and Care Excellence (NICE).** Guidance, NICE advice and quality standards [Internet]. Online. Available from: <https://www.nice.org.uk/guidance/published?ngt=Technologyappraisalguidance&ndt=Guidance>.
23. **Braun V, Clarke V.** Using thematic analysis in psychology. *Qual Res Psychol [Internet]*. 2006;**3**(2):77–101. doi: 10.1191/1478088706qp063oa.
24. **National Institute for Health and Care Excellence (NICE).** Pembrolizumab with pemetrexed and platinum chemotherapy for untreated, metastatic, non-squamous non-small-cell lung cancer. Technology appraisal guidance [TA683] [Internet]. NICE website. 2021. Available from: <https://www.nice.org.uk/guidance/ta683>.
25. **National Institute for Health and Care Excellence (NICE).** Pembrolizumab for untreated metastatic or unresectable recurrent head and neck squamous cell carcinoma. Technology appraisal guidance [TA661] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta661>.
26. **National Institute for Health and Care Excellence (NICE).** Palbociclib with fulvestrant for treating hormone receptor-positive, HER2-negative, advanced breast cancer. Technology appraisal guidance [TA619] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta619>.
27. **National Institute for Health and Care Excellence (NICE).** Obinutuzumab with bendamustine for treating follicular lymphoma after rituximab. Technology appraisal guidance [TA629] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta629>.
28. **National Institute for Health and Care Excellence (NICE).** Brentuximab vedotin in combination for untreated systemic anaplastic large cell lymphoma. Technology appraisal guidance [TA641] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta641>.
29. **National Institute for Health and Care Excellence (NICE).** Encorafenib plus cetuximab for previously treated BRAF V600E mutation-positive metastatic colorectal cancer. Technology appraisal guidance [TA668] [Internet]. NICE website. 2021. Available from: <https://www.nice.org.uk/guidance/ta668>.
30. **National Institute for Health and Care Excellence (NICE).** Nivolumab for treating recurrent or metastatic squamous cell carcinoma of the head and neck after platinum-based chemotherapy. Technology appraisal guidance [TA736] [Internet]. NICE website. 2021. Available from: <https://www.nice.org.uk/guidance/ta736>.
31. **National Institute for Health and Care Excellence (NICE).** Atezolizumab with bevacizumab for treating advanced or unresectable hepatocellular carcinoma. Technology appraisal guidance [TA666] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta666>.
32. **National Institute for Health and Care Excellence (NICE).** Isatuximab with pomalidomide and dexamethasone for treating relapsed and refractory multiple myeloma. Technology appraisal guidance [TA658] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta658>.
33. **National Institute for Health and Care Excellence (NICE).** Nivolumab for advanced non-squamous non-small-cell lung cancer after chemotherapy. Technology appraisal guidance [TA713] [Internet]. NICE website. 2021. Available from: <https://www.nice.org.uk/guidance/ta713>.
34. **National Institute for Health and Care Excellence (NICE).** Atezolizumab with carboplatin and etoposide for untreated extensive-stage small-cell lung cancer. Technology appraisal guidance [TA638] [Internet]. NICE website. 2020. Available from: <https://www.nice.org.uk/guidance/ta638>.
35. **Stensrud MJ, Hernan M.** Why test for proportional hazards? *JAMA.* 2020;**323**(14):1401–1402.
36. **Aalen OO, Cook RJ, Roysland K.** Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal.* 2015;**21**(4):579–593.
37. **Freidlin B, Korn EL.** Methods for accommodating nonproportional hazards in clinical trials: Ready for the primary analysis? *J Clin Oncol.* 2019;**37**(35):3455–3459.
38. **Marcatto F, Rolison JJ, Ferrante D.** Communicating clinical trial outcomes: Effects of presentation method on physicians' evaluations of new treatments. *Judgm Decis Mak.* 2013;**8**(1):29–33.
39. **Rutherford MJ, Lambert PC, Sweeting MJ, et al.** NICE DSU Technical support document 21. Flexible methods for survival analysis. Decis Support Unit [Internet]. 2020. Available from: www.nicedsu.org.uk.
40. **Friedlander M, Matulonis U, Gourley C, et al.** Long-term efficacy, tolerability and overall survival in patients with platinum-sensitive, recurrent high-grade serous ovarian cancer treated with maintenance olaparib capsules following response to chemotherapy. *Br J Cancer [Internet]*. 2018;**119**(9):1075–1085. doi: 10.1038/s41416-018-0271-y.