Meta-Analysis

Journal of **Comparative Effectiveness Research**

Challenges in conducting fractional polynomial and standard parametric network meta-analyses of immune checkpoint inhibitors for first-line advanced renal cell carcinoma

Svenja Petersohn[1], Bradley McGregor[2], Sven L Klijn[3], Jessica R May[3], Flavia Ejzykowicz[4], Murat Kurt[4], Matthew Dyer[3], Bill Malcolm[3], Sébastien Branchoux[5], Katharina Nickel[6], Saby George[7] & Sonja Kroep*[,1]

[1]OPEN Health Evidence & Access, Marten Meesweg 107, 3068 AV Rotterdam, The Netherlands
[2]The Lank Center for Genitourinary Oncology at Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215, USA
[3]Bristol Myers Squibb, Sanderson Rd, Denham, Uxbridge UB8 1DH, UK
[4]Bristol Myers Squibb, 100 Nassau Park Blvd #300, Princeton, NJ 08540, USA
[5]Bristol Myers Squibb, 3 Rue Joseph Monier, 92506 Rueil-Malmaison, France
[6]OPEN Health Evidence & Access, Krausenstraße 8, 10117, Berlin, Germany
[7]Roswell Park Cancer Institute, 665 Elm St, Buffalo, NY 14203, USA
*Author for correspondence: sonjakroep@openhealthgroup.com

Aim: Network meta-analyses (NMAs) increasingly feature time-varying hazards to account for non-proportional hazards between different drug classes. This paper outlines an algorithm for selecting clinically plausible fractional polynomial NMA models. **Methods:** The NMA of four immune checkpoint inhibitors (ICIs) + tyrosine kinase inhibitors (TKIs) and one TKI therapy for renal cell carcinoma (RCC) served as case study. Overall survival (OS) and progression free survival (PFS) data were reconstructed from the literature, 46 models were fitted. The algorithm entailed a-priori face validity criteria for survival and hazards, based on clinical expert input, and predictive accuracy against trial data. Selected models were compared with statistically best-fitting models. **Results:** Three valid PFS and two OS models were identified. All models overestimated PFS, the OS model featured crossing ICI + TKI versus TKI curves as per expert opinion. Conventionally selected models showed implausible survival. **Conclusion:** The selection algorithm considering face validity, predictive accuracy, and expert opinion improved the clinical plausibility of first-line RCC survival models.

Plain language summary: What was the aim of this article?: This article demonstrates an algorithm for selecting statistical models for network meta-analyses where the treatment effect varies over time. The algorithm is focused on selecting models that are closely aligned with what is seen in clinical practice.
How was this research carried out?: A network meta-analysis of four combinations of immune checkpoint inhibitors and tyrosine kinase inhibitors, and one single tyrosine kinase inhibitor for renal cell carcinoma was carried out, and 46 models were assessed. Two clinical experts were consulted separately and presented with the models to define clinical validity criteria. The plausibility was compared between models selected using expert opinion and models selected based on statistical fit.
What were the results?: The algorithm improved clinical plausibility and validity of the modelled survival, and the fit of the models with trial data. In case of overall survival, models selected purely on statistical characteristics fit poorly to the data from the start of the trial, whereas the algorithm-selected model showed better fit to the whole trial. However, challenges in the analysis of the treatments remained, since there was considerable heterogeneity across the characteristics of the included trials, which can lead to bias in the analysis.
What do these results mean?: Previous studies have not used this approach of consulting clinicians, and have solely used statistical methods for model selection. Therefore, the new model selection algorithm is a step forward compared with current practice. Further research is warranted on the evolving methods used for network meta-analyses.

Becaris PUBLISHING

First draft submitted: 17 January 2023; Accepted for publication: 27 June 2023; Published online: 11 July 2023

Keywords: health technology assessment • meta-analysis • methodology

After decades of increases, the incidence of renal cell carcinoma (RCC) has leveled off and mortality rates have decreased [1]. The latter relates partly to the rapidly evolving treatment landscape for first-line (1L) metastatic RCC. Combination treatments including immune checkpoint inhibitors (ICIs) and tyrosine kinase inhibitors (TKIs) have emerged and are displacing the former standard of care with TKI monotherapy such as sunitinib (SUN) [2]. Four different combinations of VEGF and ICI are currently approved by the US FDA for management of RCC [3–6]. For reimbursement decision-making, indirect treatment comparisons such as network meta-analysis (NMA) are frequently used to compare outcomes of 1L RCC treatment options that have not been compared in head-to-head clinical trials [7–10], and to quantitatively synthesize all available evidence into pooled efficacy and safety outcomes that can inform health-economic evaluations [11]. When comparing different regimens, hazard ratio-based NMAs for relative differences in overall survival (OS) and progression-free survival (PFS) have often been based on Cox proportional hazards (PH), as the treatment effect was assumed to be constant over time (i.e., the hazard ratio [HR] between the treatments was constant) [12].

Targeting different receptors, the mechanisms of action (MoAs) of currently approved ICI + TKI combination therapies based on avelumab in combination with axitinib (AVE + AXI), nivolumab plus cabozantinib (NIVO + CABO) and pembrolizumab (PEM) in combination with AXI or with lenvatinib (LEN), are considerably different and offer improved outcomes compared with TKI monotherapy [2]. Given the proportional hazards assumption not holding in the trials comparing ICI + TKI and TKI treatments, HR-based NMAs would not be adequate for the synthesis of these trials, and an alternative NMA approach relaxing the PH assumption should be used [13].

Such alternative approaches relaxing the PH assumption are NMAs using time-varying relative treatment effects by fitting standard parametric models or fractional polynomials (FP) [14–16]. Of these, the FP method may be a more popular method, also used in National Institute for Health and Care Excellence and Haute Autorité de Santé technology appraisals [13]. However, recent NMAs using FP models have been deemed uncertain as the analysis results obtained suffered from face validity issues relating to clinically implausible extrapolations of survival [17,18].

Potential reasons for face validity issues could lie in the similarity assumption underlying all NMAs being violated: trials should be sufficiently similar in terms of effect modifiers. Effect modifiers are population characteristics such as age, disease severity and risk grouping that might influence the relative treatment effect in a trial. These must be comparable for all trials included in the NMA to ensure an unbiased estimate [11]. Differences between trials can bias the relative treatment effect estimated by the NMA; additionally, they can complicate the estimation of absolute survival outcomes (i.e., survival curves), as different trials may provide different estimates of TKI effectiveness to which the ICI + TKI HR is applied to estimate absolute outcomes. Researchers must assess these differences to choose a study or pooled studies presenting baseline survival that is representative of the decision problem.

Next to this, methodological features specific to the time-varying NMA approaches may play a role in the estimation of results lacking face validity: to deal with non-proportional relative treatment effects over time, the FP NMA specifically fits continuous functions of different shapes with one or two parameters to model the relative treatment effect over time (i.e., first-order and second-order models) and uses different power combinations for the parameters [19]. The approach using standard parametric models fits five standard parametric models to the trial data (exponential, Gompertz, Weibull, log-normal, log-logistic) [16]. The use of time-varying hazard NMA models thus requires that a realistic model is selected from the pool of available models, which can contain as many as 49 models, eight first-order FP models, 36 second-order FP models, and five standard parametric models. The selected model should fit well with the observed trial data, but also provide extrapolation estimates that are in line with clinical expectations and thus meet plausibility and face validity criteria.

However, model selection is usually exclusively based on statistical fit criteria rather than clinical plausibility, emphasizing the fit of the model with observed data from the trial period [15]. Model selection based on statistical fit criteria, often using the deviance information criterion (DIC), considers the overall fit of the time-varying hazard model to the data of all trials. Therefore, difficulties with preferential fitting to some trials can arise, driven by their larger sample size, longer follow-up time, or their heterogeneity of hazard pattern. Additionally, overfitting and clinically implausible extrapolations past the observed trial period may go unnoticed, as this is not assessed by the

DIC. The use of DIC may thus enable the dismissal of time-varying hazards models altogether, as the statistically best-fitting time-varying hazards model may lack face validity, and arbitrary selection of another model without explicit criteria would lack transparency and justification.

With FP and standard parametric NMAs becoming increasingly relevant to HTA authorities and utilized in the RCC and wider oncology setting [14,20,21], a transparent and clinically plausible model selection approach is needed. The purpose of this paper is to outline a new structured approach for the selection of FP and standard parametric NMA models that can be applied specifically in the analysis of ICI + TKI versus TKI monotherapy in 1L RCC but also more widely for other NMAs of novel oncology therapies. We describe the new model selection approach based on face validity and clinical plausibility of the survival curves, and predictive accuracy against trial data, and we compare the resulting PFS and OS outcomes to outcomes obtained from model selection based on statistical fit criteria alone.

Materials & methods

A time-varying hazards NMA including FP and standard parametric models was conducted, followed by model selection based on statistical fit criteria and model selection based on an algorithm considering statistical fit and clinical plausibility, see McGregor BA, Petersohn S, Klijn S *et al.* [22] for further detail. The analysis was conducted using the following sequential steps: feasibility assessment of the NMA evaluating similarity assumption and PH assumption, trial data preparation, time-varying hazard model fitting, model selection and comparison of selected models based on conventional and algorithm approach.

Feasibility assessment & data preparation

A systematic literature review [23] and updated search for publications up to December 2021 identified clinical trials reporting survival outcomes of ICI + TKI combinations in 1L advanced RCC in the all-risk (i.e., intention to treat) population as defined by the International Metastatic RCC Database Consortium (IMDC) score [24]. For these trials, networks of evidence were constructed for PFS and OS during the feasibility assessment phase, after which heterogeneity among included trials was assessed according to pre-specified criteria (Supplementary Section 1.1.1). As individual patient-level data (IPD) from three relevant trials was unavailable, Kaplan–Meier (KM) curves of these trials were digitized using the WebPlotDigitizer [25] and pseudo-IPD was reconstructed using the methodology from Guyot [26].

The PH assumption was tested by visual assessment of the log-cumulative hazards and Schoenfeld residual plot, in addition to the Grambsch Therneau test (using available IPD from CheckMate 9ER trial [NCT03141177] and pseudo-IPD of all other trials [27–29]; see Supplementary Section 1.2.1 for assessment criteria). After between-trial heterogeneity had been assessed and the non-proportionality of hazards had been tested, trial data to be used in the time-varying hazard NMA was prepared. The (pseudo)-IPD were formatted for FP model fitting as stated in Jansen 2011 [15] and plotted for visual inspection.

Time-varying hazard model fitting

The network of evidence highlighted that for each comparison between treatments only one trial was available, CLEAR comparing PEM + LEN and SUN, CheckMate 9ER comparing NIVO + CABO and SUN, JAVELIN Renal 101 comparing AVE + AXI and SUN, and KEYNOTE-426 comparing PEM + AXI and SUN, see (Supplementary Figure 1) [27–30]. Furthermore, no prior beliefs on the between-study heterogeneity were available. Fixed-effects models were therefore fitted, although random effect models were explored using vague priors to inform between-study heterogeneity, in line with published guidelines [31]. The analysis was conducted in R [32] and executed using the software package Stan [33]. In line with the approach described by Jansen [15], FP NMA models were fitted with one or two parameters of different powers; these were selected from the following set: -2, -1, -0.5, 0, 0.5, 1, 2, 3. This resulted in eight first-order and 36 second-order models. Additionally, four of the five standard parametric models were fitted (the exponential model was left out as it represented a PH model), of which the Weibull and Gompertz models overlapped with first-order FP models and were considered as such, leaving two unique models (the log-logistic and log-normal models).

After the models were fitted, the relative treatment effect over time within each trial was estimated for each time-varying HR model. The absolute survival (i.e., PFS and OS) were estimated by applying the relative treatment effects to the anchor treatment arm. In this evidence network, SUN served as the anchor treatment, connecting all other ICI + TKI treatments. The similarity of the trials' SUN curves was assessed by comparing confidence

intervals and point estimates at different time points. For the PFS network, the SUN arm of KEYNOTE-426 differed from the remaining trials (Supplementary Table 3). SUN arms from all included trials were pooled and used as the anchor, as there was no clear indication that one trial was more reflective of clinical reality than others, to counter the impact of observed differences between trials regarding several population characteristics on the PFS_{SUN} curve. For the OS network, the SUN arms of CheckMate 9ER and JAVELIN differed from those of CLEAR and KEYNOTE-426 (Supplementary Table 6). The SUN arm of the most recent trial at time of the initial analysis (CheckMate 9ER) was selected as anchor to reflect current clinical practice in terms of subsequent treatment post 1L SUN. The subsequent ICI treatment option is of evident importance, as SUN first-line treatments will show higher survival over time when patients can switch to ICI treatments, compared with older trials where such an option was not available or subsequent ICI use was much lower.

Model selection approaches

Converging models identified by an R-hat <1.2 were considered in the model selection [34]. Model convergence, sign of the model successfully fitting to the trial data, was assessed utilizing trace plots, the autocorrelation function, Monte Carlo error size, and the Kernel density plots. This is in line with the definition of convergence in a Bayesian framework implying how a measured outcome limits toward a stationary distribution that remains the same as time progresses [35].

Model selection based on conventional statistical fit criteria used solely the DIC. For the clinically focused algorithm, two clinical experts' opinions were elicited separately. Experts were presented with survival extrapolations, HRs, hazard curves, and an overview of modeled versus observed survival outcomes for viable models. Expert opinions on the plausibility of the modeled ranking of treatments and long-term survival expectations were also elicited. Consequently, the clinically focused algorithm includes face validity checks to prevent the selection of models with overfitting issues (i.e., when extreme values or outliers in HRs are captured by the model rather than the HR trend), implausible trends over time, counterintuitive ordering of treatments in terms of survival or hazards, and extrapolated outcomes that do not reflect external long-term evidence clinical expert opinion. The model selection algorithm steps aligned with Figure 1 were:

1. Face validity of first-order and second-order models: rank converged models by DIC. Before anchoring, does the model visually fit the trial data (KM curves)? Do first-order models appear too inflexible, or do second-order models show too much flexibility?
2. Goodness of fit: Are modeled median survival, absolute survival at the latest available landmark (24 months), and restricted mean survival within the 95% confidence interval of the underlying observed trial outcomes? Do modeled hazards align with trial hazards after anchoring?
3. Clinical plausibility: Do HR, hazard curves, and survival extrapolation curves reflect clinical expectations – are the ranking of treatments in terms of predicted hazards and survival, and trends over time, as expected? Is the length of survival and the proportion of long-term survivors reasonable? (Experts were first asked to express their expectations regarding treatment rankings, survival trends over time and expectation of survival proportions at several landmarks, then figures comparable to Figures 2 & 3 of models meeting those criteria were made available to aid further discussion of the criteria).
5. Consider DIC for the final model selection in case multiple viable models were identified (based on steps 2 and 3).

Finally, selected models based on DIC alone and selected based on the algorithm were compared based on predictive accuracy and clinical plausibility of survival extrapolations.

Results

Feasibility assessment & data preparation

Networks of evidence were generated for PFS and OS using the CheckMate 9ER, CLEAR, JAVELIN Renal 101, and KEYNOTE-426 trials [27,28,36], including the treatments NIVO + CABO, PEM + LEN, AVE + AXI, PEM + AXI, anchored by SUN as the common comparator used in all trials (Supplementary Section 1.3 for network diagram). The similarity assessment highlighted the presence of considerable heterogeneity for several potential treatment effect modifiers, most evidently for Eastern Cooperative Oncology Group (ECOG) performance score,

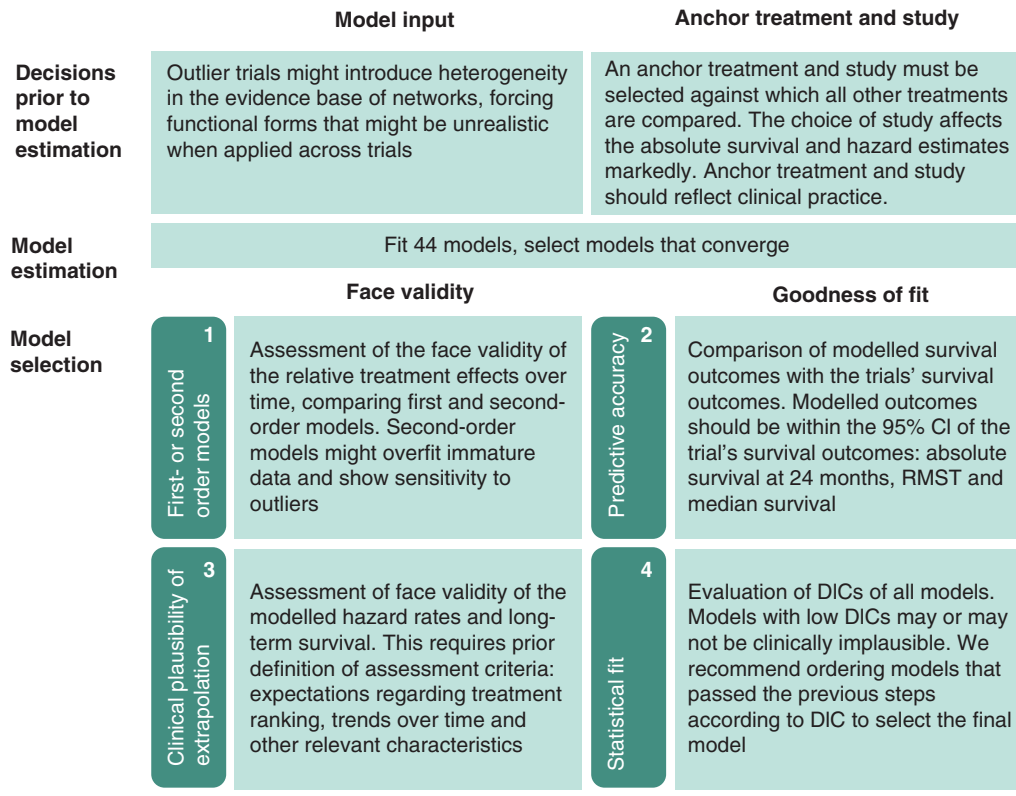

Figure 1. Model selection algorithm.
CI: Confidence interval; DIC: Deviance information criterion; RMST: Restricted mean survival time.

IMDC risk scores, prior nephrectomy, prior radiation therapy, PD-L1 expression, and number of metastatic sites (Supplementary Section 1.1.2).

The assessment of the PH assumption showed conflicting results for the PFS and OS networks. In the PFS network, the PH assumption was rejected for the CheckMate 9ER and CLEAR trials. In the OS network, the PH assumption was rejected overall, as the KEYNOTE-426, CheckMate 9ER, and CLEAR trials violated the assumption, and mixed results were seen for the JAVELIN Renal 101 trial (Supplementary Section 1.2.2). A time-varying hazard NMA was thus considered appropriate for OS and PFS. The input data used for the analyses is presented in the Appendix. It visualizes the differences between SUN arms in the PFS network and shows different patterns of crossing and non-crossing between SUN and ICI + TKI arms in the OS network of evidence, both likely related to heterogeneity.

Time-varying hazard models

Random effects models did not converge. Of the 44 fixed effects FP models and two standard parametric models fitted, 37 models converged for PFS and 29 models converged for OS. Non-converging models were first- and second-order models with p1 = 3 and p1 = 2, possibly as the flexibility of the models was not supported by a sufficient evidence base. Models with the best statistical fit were second-order models p1 = -2, p2 = -2 for PFS, and p1 = -2, p2 = 1 for OS.

Model selection algorithm-based approach

All 37 and 29 converged first- and second-order models for PFS and OS, respectively, were assessed for the plausibility of their HRs versus SUN over time (Figure 2). For PFS, first- and second-order models were generally considered viable. For OS, second-order models were excluded as these modeled the underlying data with too much flexibility resulting in extremely variable and implausible time-varying HRs, mostly for the PEM + LEN arm (Supplementary Section 2.2.1).

After anchoring the relative effects of PFS to the pooled SUN arms of all trials and anchoring the relative effects of OS to the SUN arm of the CheckMate 9ER trial, survival outcomes were assessed for predictive accuracy against observed outcomes of the underlying trials. For PFS, anchoring to the pooled SUN arm of all trials increased the survival outcomes (absolute survival at 2 years, median survival, restricted mean survival) of NIVO + CABO, PEM + LEN, and AVE + AXI compared with the trial data, and decreased the survival outcomes of PEM + AXI to some extent (i.e., estimated 2-year survival was 3–5% lower than observed in the trial; see the Supplementary Section 2.3.1). This was in line with the differences seen between the SUN arms of the trials, and the bias introduced by anchoring PFS to a pooled rather than one individual trial's SUN arm. Most PFS models thus showed somewhat inadequate predictive accuracy against some of the available trial data, but none of the models were excluded. All OS models passed the assessment of predictive accuracy against trial data; all modeled versus trial outcomes for OS and PFS are provided in the (Supplementary Section 2.3.2).

As a part of the final selection step, the models were visually inspected for clinical plausibility of their survival extrapolations. For PFS, 34 models were excluded from consideration due to lack of clinical plausibility as hazards and survival curves showed unlikely rankings of treatments and clinically implausible optimistic extrapolation outcomes (e.g., estimating SUN PFS plateauing from 3 years at 13%), leaving three viable models. For OS, six models were excluded from consideration based on clinical long-term survival expectations, resulting in two viable models. The final models, first-order PFS model with p1 = -2 and first-order OS model with p1 = 0, were selected based on DIC, as all remaining models were considered clinically plausible.

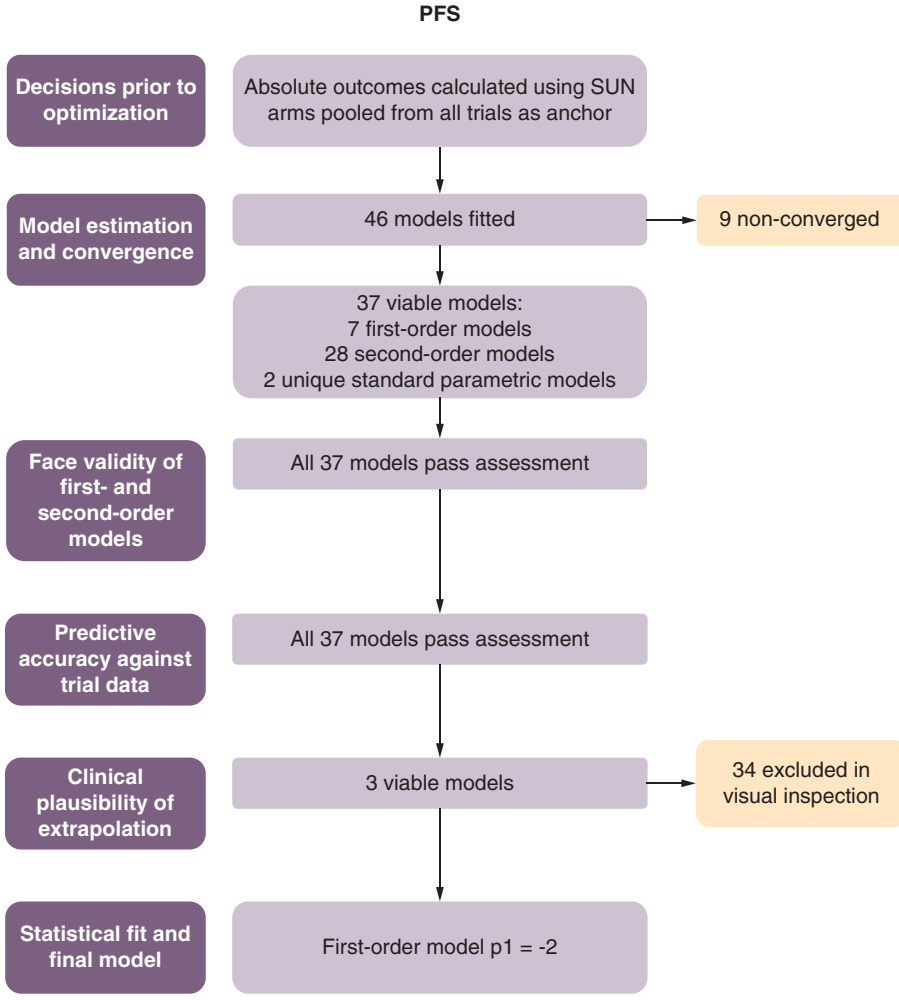

Figure 2. Model selection algorithm progression-free survival and overall survival.
OS: Overall survival; PFS: Progression-free survival.

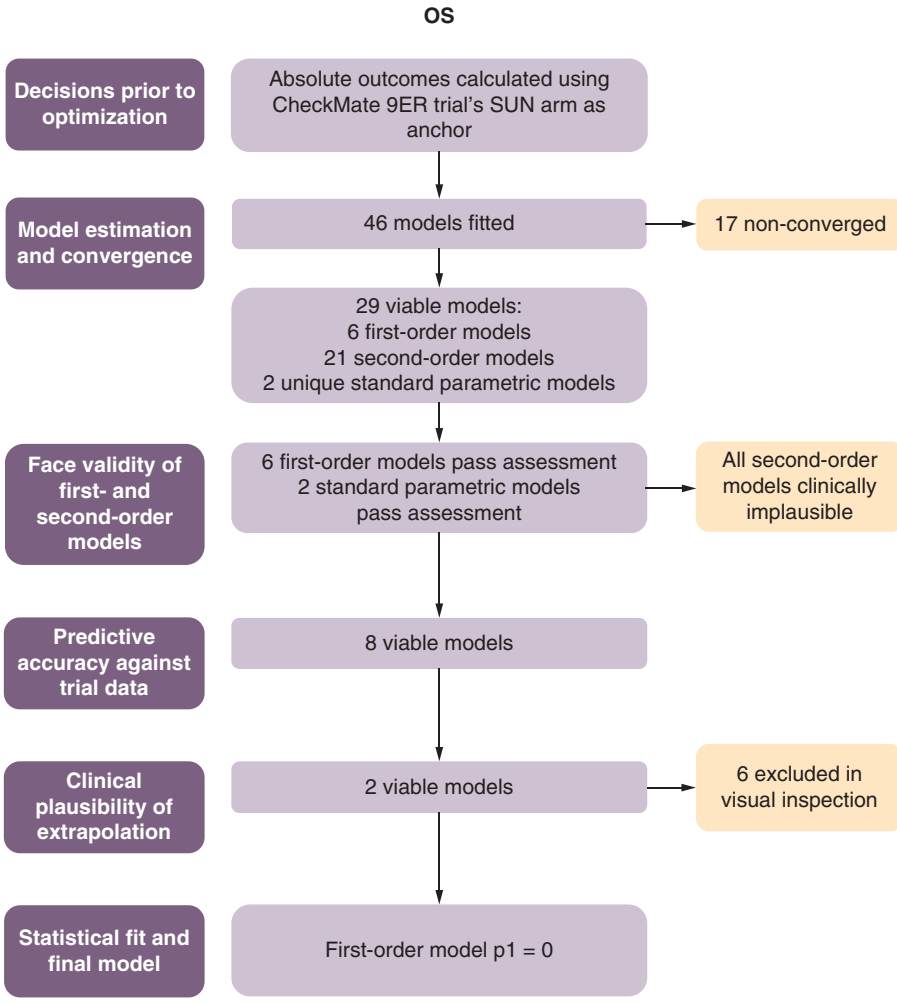

Figure 2. Model selection algorithm progression-free survival and overall survival (cont.).
OS: Overall survival; PFS: Progression-free survival.

Survival extrapolations

Progression-free survival

Both the algorithm-selected and statistical fit-based PFS models provided adequate visual fit with the trial data, with the statistical fit-based model showing better fit before anchoring for all treatments except PEM + LEN (Figure 3A), despite featuring visually different hazard profiles (Figure 3B). The algorithm-based model showed a better fit after anchoring for all trials except for PEM + AXI, as reflected in the survival outcomes presented in Table 1. While predicting similar long-term outcomes, the resulting survival curves of the statistical fit-based selected model showed a crossing of AVE + AXI and PEM + AXI which was not produced by the algorithm-based selected model (Figure 3C). Both models predicted the highest PFS for PEM + LEN and NIVO + CABO and the lowest PFS for SUN.

Overall survival

The DIC-selected OS model fitted poorly to the initial months of the PEM + LEN arm (Figure 4A) and estimated exceedingly high hazards at onset for PEM + LEN and AVE + AXI and after 30 months for PEM + LEN. Comparing survival outcomes after anchoring (Table 1), the algorithm-selected model showed a better fit overall, likely related to the modeling of crossing between SUN and ICI + TKI arms which occurred earlier in the DIC-based model than in the algorithm-based model (Figure 4A). Apart from PEM + LEN and AVE + AXI, the models predicted similar absolute survival outcomes at 2 years, but vastly different median survival and absolute survival at 5 years. The long-term extrapolations of the models showed the DIC-based model predicting extremely

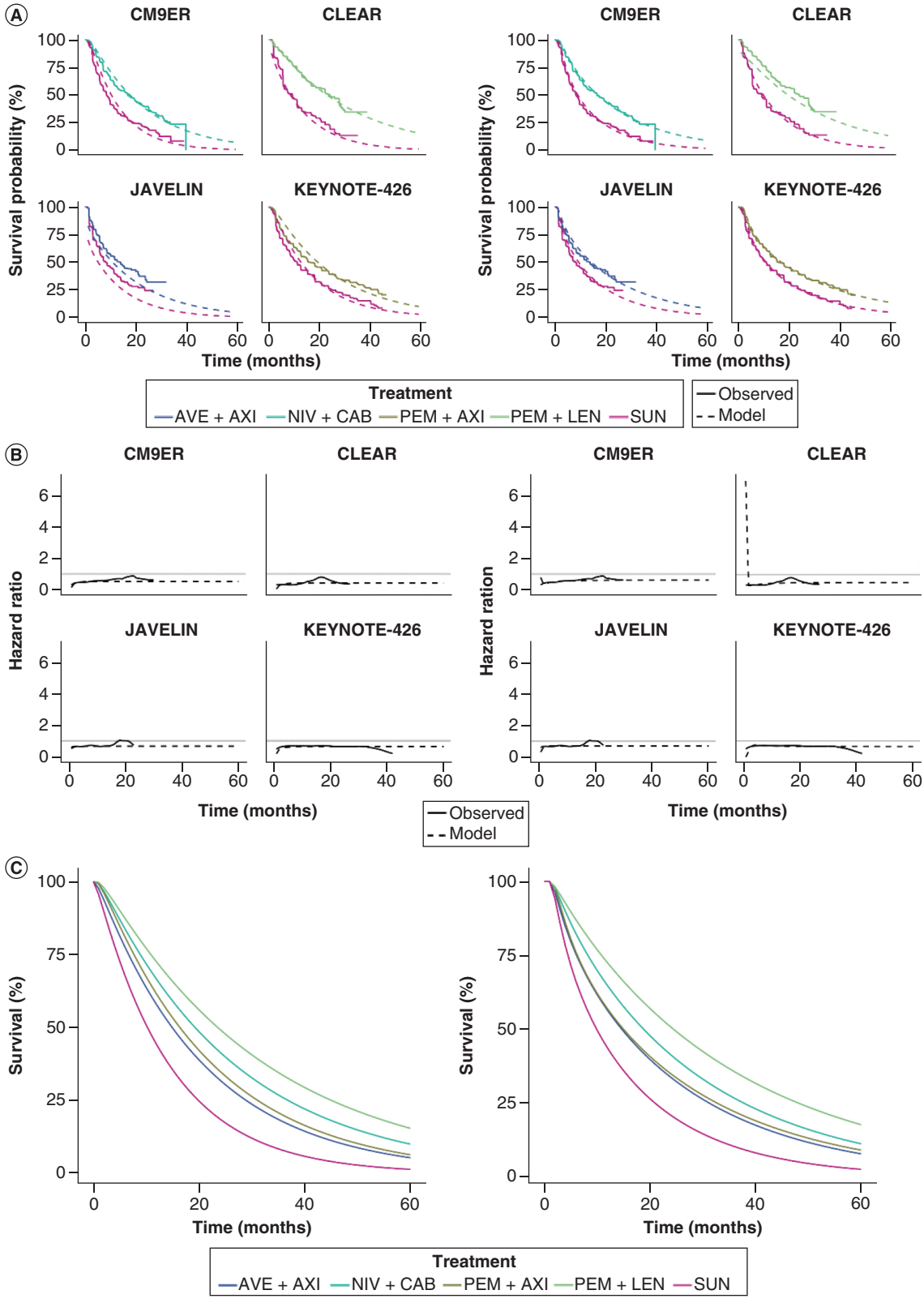

Figure 3. Survival outcomes of the progression-free survival models selected by (left) algorithm model p = -2, and (right) DIC p1 = -2, p2 = -2. (A) Modeled before anchoring versus trial survival curves, **(B)** modeled hazard ratios (HRs) and trial HRs versus SUN, **(C)** survival curves after anchoring.
AVE: Avelumab; AXI: axitinib; CAB: Cabozantinib; LEN: Lenvatinib; NIV: Nivolumab; PEM: Pembrolizumab; SUN: Sunitinib.

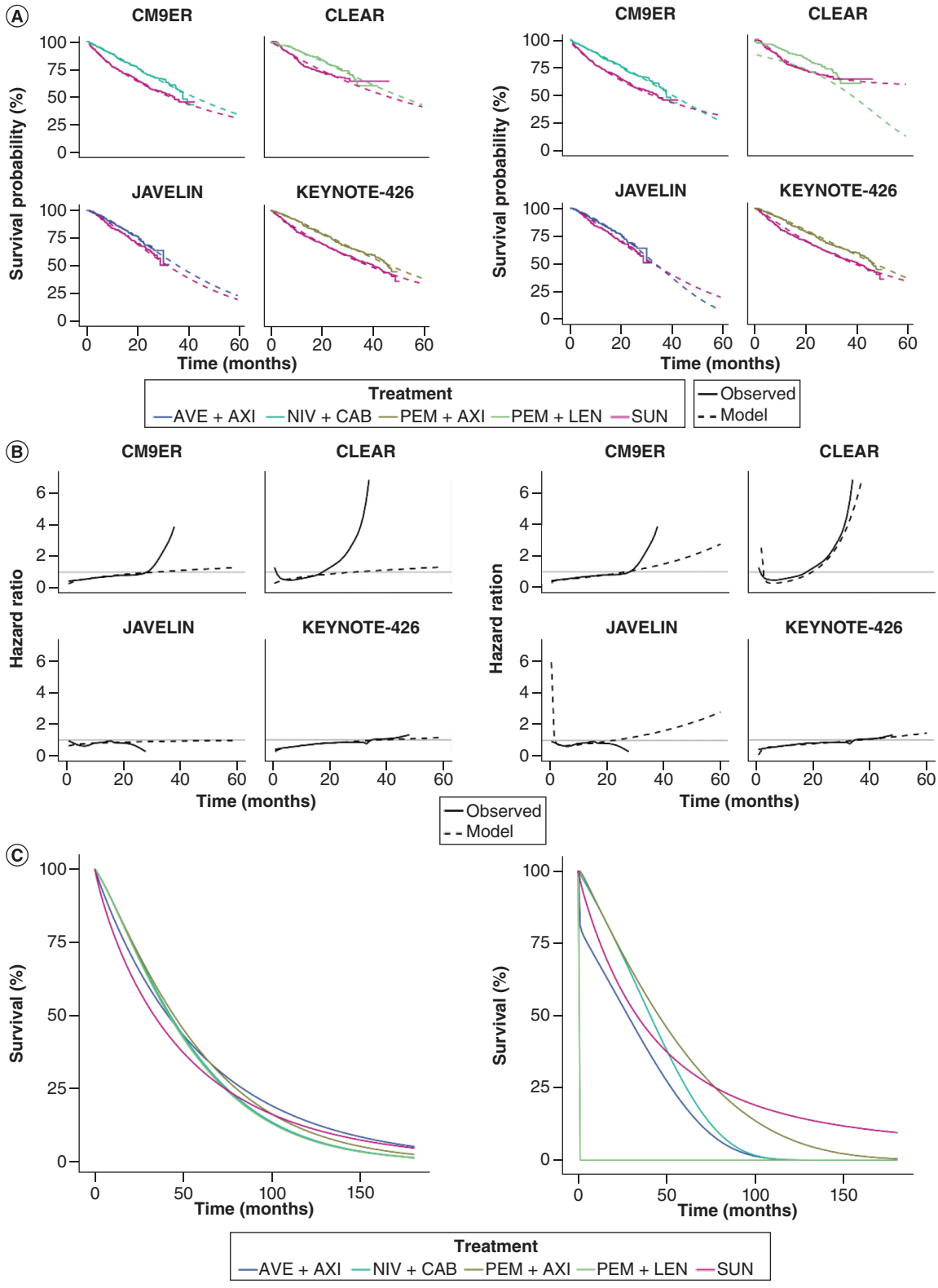

Figure 4. Survival outcomes of the overall survival models selected by (left) algorithm model p = 0, and (right) DIC p1 = -2, p2 = 1. (A) Modeled before anchoring versus trial survival curves, **(B)** modeled HRs and trial HRs versus SUN, **(C)** survival curves after anchoring.
AVE: Avelumab; AXI: axitinib; CAB: Cabozantinib; LEN: Lenvatinib; NIV: Nivolumab; PEM: Pembrolizumab; SUN: Sunitinib.

10.57264/cer-2023-0004

Table 1. Comparison of survival outcomes between trials and NMA models.

	Trial	Algorithm-based model	DIC-based model
PFS		**p1 = -2**	**p1 = -2, p2 = -2**
Survival at 2 years:			
PEM + LEN	46.1% (40.1%, 53.1%)	49.2% (39.9%, 57.1%)	50.7% (29.9%, 59.6%)
NIVO + CABO	38.8% (33.6%, 44.9%)	41.5% (30.8%, 49.3%)	41.3% (31.0%, 50.2%)
AVE + AXI	29.1% (22.3%, 37.9%)	31.9% (22.7%, 40.1%)	33.7% (24.4%, 43.3%)
PEM + AXI	38.3% (33.7%, 43.5%)	34.9% (26.6%, 41.6%)	34.9% (27.1%, 42.9%)
Survival at 5 years:			
PEM + LEN	–	15.4% (9.3%, 22.9%)	17.5% (7.1%, 27.2%)
NIVO + CABO		10.1% (5.5%, 15.4%)	11.2% (5.7%, 18.4%)
AVE + AXI		5.3% (2.4%, 9.5%)	7.8% (3.3%, 14.9%)
PEM + AXI		6.4% (3.6%, 10.0%)	9.1% (5.0%, 15.0%)
Median survival (months):			
PEM + LEN	22.8 (19.7, 26.3)	23.8 (18.8, 28.7)	24.2 (10.6, 30.6)
NIVO + CABO	16.6 (13.0, 19.9)	19.4 (13.7, 23.0)	19.1 (13.9, 23.1)
AVE + AXI	12.3 (10.1, 15.1)	15.0 (11.1, 17.8)	15.1 (11.2, 18.4)
PEM + AXI	15.5 (13.9, 19.8)	16.7 (12.5, 19.3)	15.5 (11.8, 18.4)
OS		**p1 = 0**	**p1 = -2, p2 = 1**
Survival at 2 years:			
PEM + LEN	77.5% (72.8%, 82.5%)	70.5% (58.2%, 79.7%)	0.0%† (0.0%, 14.8%)
SUN	60.2% (55.0%, 65.8%)	59.7% (52.8%, 66.0%)	59.1% (38.7%, 67.7%)
NIVO + CABO	70.1% (65.3%, 75.4%)	71.1% (65.2%, 76.6%)	71.8% (50.3%, 78.2%)
AVE + AXI	64.1% (57.5%, 71.4%)	66.5% (52.3%, 76.5%)	54.2% (0.0%, 78.0%)
PEM + AXI	74.5% (70.5%, 78.8%)	71.7% (62.5%, 78.8%)	72.4% (40.5%, 80.0%)
Survival at 5 years:			
PEM + LEN	–	33.9% (14.1%, 55.0%)	0.0%† (0.0%, 1.3%)
SUN		31.3% (22.2%, 41.1%)	32.0% (12.9%, 47.7%)
NIVO + CABO		34.7% (24.4%, 45.5%)	26.1% (7.7%, 44.8%)
AVE + AXI		36.7% (15.9%, 56.8%)	18.6% (0.0%, 61.6%)
PEM + AXI		37.3% (22.9%, 51.6%)	37.2% (9.8%, 58.0%)
Median survival (months):			
PEM + LEN	NA (32.6, NA)	41.9 (31.7, 62.2)	0.0† (0.0, 0.0)
SUN	33.0 (29.0, NA)	33.1 (27.6, 40.1)	32.9 (13.3, 52.2)
NIVO + CABO	36.1 (35.5, NA)	43.0 (37.1, 51.1)	40.4 (24.6, 53.5)
AVE + AXI	29.5 (26.0, NA)	40.5 (29.4, 62.2)	27.7 (0.0, NA)
PEM + AXI	44.2 (42.7, 47.5)	44.9 (35.8, 58.0)	42.8 (0.0, 67.1)

†DIC-based model fit on overall data resulted in a highly implausible OS curve for PEM + LEN.
For SUN, progression-free survival was pooled across trials, hence no comparison to a single trial estimate is presented.
AVE: Avelumab; AXI: axitinib; CAB: Cabozantinib; DIC: Deviance information criterion; LEN: Lenvatinib; NIV: Nivolumab; OS: Overall survival; PEM: Pembrolizumab; PFS: Progression free survival; SUN: Sunitinib.

low survival outcomes for AVE + AXI and LEN + PEM (19% and 0%, respectively, vs 32% with SUN), which may lack face validity.

Discussion

The objective of this study was to develop an algorithm that improves the time-varying hazard NMA model selection for predictive accuracy, face validity, expert opinion regarding clinical plausibility, and goodness of fit using a case study of ICI + TKI versus TKI monotherapy in 1L RCC. While the developed algorithm improved the clinical plausibility of the results, the validity of the survival extrapolations and accuracy of the model extrapolations compared with the observed survival was still found to be inadequate, especially for OS.

Potential reasons for the remaining issues likely relate to heterogeneity and can be further differentiated into three problems: violations of the homogeneity assumption and resulting overfitting, biased relative effectiveness outcomes and anchoring to heterogeneous SUN survival data. First, heterogeneity in the OS network of evidence, as highlighted by different degrees of crossing of the SUN and ICI + TKI arms, led to overfitting of more flexible second-order OS models to the data. While time-varying hazard models relax the PH assumption, one functional form is fitted to all trials in line with the assumption of similarity, homogeneity, and consistency of the underlying trials. A violation of the assumption would cause for biased outcomes, as is the case where there are different MoAs that potentially show different hazard trends over time that cannot be feasibly modeled with one functional form, as for ICI + TKI versus TKI treatments. Thus, this assumption of homogeneity and the appropriateness of one

functional form can be questioned, particularly for the OS data where some curves cross, and the outcomes should therefore be interpreted with caution.

Second, there was also a high level of observed and unobserved heterogeneity between trials with clinical factors such as observed baseline ECOG performance score and IMDC risk classifications varying among trials (Supplementary Section 1.1.2), and potentially unobserved intertumor and intratumor heterogeneity (e.g., due to genetic or tumor microenvironment differences) [37]. This observed heterogeneity, as well as suspected unobserved heterogeneity, challenges the homogeneity assumption of the NMA and may lead to biased results [37,38]. This was observed in a recently published matched-adjusted indirect comparison of the CheckMate 9ER and KEYNOTE-426 trials, where in the matched population adjusted for known prognostic factors, median PFS of NIVO + CABO increased from 17.0 to 19.3 months [39]. Accounting for differences in relevant effect modifiers and population characteristics may thus play a central role in the indirect comparison of 1L RCC trials, and covariate adjustment may be needed [40]. Additionally, the variation in OS curves crossing may also be reflective of differences in subsequent treatment in the different ICI + TKI and SUN arms. Highly effective subsequent treatment (e.g., ICI, ICI + TKI or ICI + ICI treatment) can be given after progression on SUN [41], which biases and increases OS in the SUN arm of some trials. Use of subsequent ICI treatment might have been twice as high in CLEAR and KEYNOTE-426 SUN arms compared with JAVELIN renal 101 and CheckMate 9ER SUN arms [27–30]. The degree to which such highly effective subsequent treatment is used is likely affected by geographical setting of the trials. If feasible given the evidence base, advanced methods, such as covariate adjustment, could be applied to mitigate the effect. In other settings, adjustment for crossover may be appropriate, if the treatments are not licensed in subsequent lines, and assuming that IPD is available.

Third, PFS differences between the SUN arms caused discrepancies between modeled outcomes and trial outcomes for ICI + TKIs. This is due to the computation of absolute outcomes being driven by SUN defined as the anchor against which all other treatments are compared. Different anchors, in this case different SUN arms, can be chosen from the network, and these will lead to different absolute survival outcomes for ICI + TKIs. In the presence of heterogeneous PFS between SUN arms, absolute survival outcomes need to be interpreted in the context of the selected anchor arm, as the anchoring to one of the SUN arms will inevitably introduce discrepancies for those ICI + TKIs for which the comparator SUN arm performed differently. It must be assumed that all issues related to heterogeneity, violations of the homogeneity assumption resulting in overfitting, biased relative effectiveness outcomes, and anchoring issues with heterogeneous survival data, are likely generalizable to other indications where treatments with different MoAs are used.

Regarding the newly-developed model selection algorithm, exchange with clinical experts for the definition of face validity and plausibility criteria highlighted that due to OS data immaturity in the trials and the novelty of the ICI + TKI treatments, it is not certain if and how long initial survival benefits of ICI + TKI treatments hold over time, or if TKI treatment followed by ICI-containing subsequent treatment will result in better outcomes given the increase in HR for OS with time across all TKI + ICI trials. This limitation cannot be overcome by the use of the model selection algorithm, as expert elicitation is being conducted on long-term estimates that are currently not observable in practice. This could introduce subjectivity and bias, and future expert elicitation for the purpose of model selection may benefit from a structured process including a larger sample of experts, as discussed in other contexts [42]. A further limitation inherent to the selection of one model from a pool of available models that is not resolved by the use of the selection algorithm is preferential selection. Although the described algorithm aims to omit the possibility of choosing the most favorable model by determining rules for survival extrapolation plausibility *a priori*, further steps toward transparency can be taken. This may include the selection of a curve aligning with available evidence including real-world data and clinical beliefs on underlying disease progression and treatment effect before the NMA is conducted, rather than based on trial data, followed by fitting of the curve to the trials included in the analysis [17].

The results of the analysis indicate that despite the use of advanced model selection methods for time-varying models in 1L RCC, limitations persist relating to the presence of heterogeneity in the trials included in the analysis, as opposed to previous analyses in other indications [13]. A number of studies have been published comparing therapies in RCC using NMAs [7–10]; however, only few of these studies considered the non-proportionality of hazards, which is likely to result in misleading or biased conclusions on the relative efficacy of RCC treatments. For the studies that did use a time-varying hazard approach, model selection was solely based on statistical fit, and long-term extrapolations were not evaluated as the NMA outputs were not part of a health-economic evaluation [43]. This is also seen in publications in other oncology indications, where relative effects comparisons were provided without

considering anchored absolute outcomes, thus the input to health-economic evaluations was not considered [44]. As such, the method for the selection of time-varying NMA models applied in our study is a step forward compared with available literature, and we encourage researchers in other tumor areas to apply and, where possible, refine our decision algorithm approach further to evaluate whether these findings are consistent across other tumor types and novel oncology therapies. However, limitations and challenges relating to the use of NMAs in the presence of heterogeneity are yet to be resolved. Given the challenges observed in this case study using FP and standard parametric models, evolving new methods for non-proportional hazards NMA may offer advantages over the models used in this analysis [13], warranting further research and supporting the use of similar structured selection criteria when assessing the feasibility and performance of different time-varying hazards NMA methodologies [45].

Conclusion

This study presented an algorithm improving model selection for the time-varying hazards NMAs by considering face validity, predictive accuracy, and expert opinion. The developed algorithm improved clinical plausibility of the results; however, the high level of heterogeneity present in advanced/metastatic RCC trials regarding patient characteristics, subsequent treatment and differences in MoAs hindered an unbiased indirect treatment comparison of survival outcomes of ICI + TKI combinations.

Summary points

- Time-varying hazard models are commonly used for quantitative evidence synthesis in a network meta-analysis (NMA) framework for Health technology assessment-decision making.
- In first-line renal cell carcinoma (RCC), time-varying hazards NMA is recommended as the treatment effect of immune checkpoint inhibitors + tyrosine kinase inhibitors (TKI) versus TKI monotherapy varies over time.
- Time-varying NMA model selection considering clinical plausibility over statistical fit improves the validity of survival estimates.
- The elicitation of expert opinion for NMA model selection helped identify important assumptions underlying the extrapolation models, and areas of uncertainty.
- Structured expert elicitation approaches may reduce the possibility of subjectivity and bias of expert opinion.
- In this case study in RCC, heterogeneity and treatment effect modifiers hampered the validity of survival extrapolations.
- The suitability of non-proportional hazards NMA can be called into question for the comparison of treatments with different mechanisms of action, given the underlying violations of the homogeneity assumption.
- Tailored time-varying NMA methods are needed for indications or therapy comparisons where established approaches fail.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: https://bpl-prod.literatumonline.com/doi/10.57264/cer-2023-0004

Author contributions

All authors contributed to the conceptualization, reviewing and editing of the publication. The analyses were conducted by S Petersohn, K Nickel and S Kroep.

Acknowledgments

The authors thank John Borrill for his comments on the manuscript, and Martijn Simons for supporting the analysis.

Financial & competing interests disclosure

The study was funded by Bristol Myers Squibb. The funder contributed to the study design, collection, analysis, and interpretation of the data in collaboration with the authors of this manuscript. SL Klijn, JR May, F Ejzykowicz, M Kurt, M Dyer, B Malcolm and S Branchoux are employed by Bristol Myers Squibb (BMS). S Petersohn, K Nickel and S Kroep are employed by OPEN Health Company and are consultants for BMS. S George received grants from or advised Pfizer, Merck, Agensys, Novartis, Bristol Myers Squibb, Bayer, Eisai, Seattle Genetics/Astellas, Calithera Biosciences, Corvus Pharmaceuticals, Surface Oncology, Exelixis, Aravive, Aveo, EMD Serono, QED Therapeutics, Sanofi/Genzyme and Gilead Sciences over the past 3 years outside of the submitted work. BMG received grants from or advised Astella, Bayer, Bristol Myers Squibb, Calithera, Dendreon, Exelixis, Ipsen, Pfizer, Seattle Genetics,

Aptitude Health, MJH, Targeted Oncology, OncLive, DAVA Oncology and Curio. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Data sharing statement

The authors certify that this manuscript reports the secondary analysis of clinical trial data that have been shared with them, and that the use of this shared data is in accordance with the terms (if any) agreed upon their receipt. The source of this data is: NCT03141177.

Open access

References

1. Escudier B, Porta C, Schmidinger M *et al.* Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-updagger. *Ann. Oncol.* 30(5), 706–720 (2019).
2. Rassy E, Flippot R, Albiges L. Tyrosine kinase inhibitors and immunotherapy combinations in renal cell carcinoma. *Ther. Adv. Med. Oncol.* 12, DOI: 10.1177/1758835920907504 (2020).
3. FDA. FDA approves avelumab plus axitinib for renal cell carcinoma (2019). www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-avelumab-plus-axitinib-renal-cell-carcinoma
4. FDA. FDA approves nivolumab plus cabozantinib for advanced renal cell carcinoma (2021). www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-nivolumab-plus-cabozantinib-advanced-renal-cell-carcinoma
5. FDA. FDA approves pembrolizumab plus axitinib for advanced renal cell carcinoma (2019). www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-plus-axitinib-advanced-renal-cell-carcinoma
6. FDA. FDA approves lenvatinib plus pembrolizumab for advanced renal cell carcinoma (2021). www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-lenvatinib-plus-pembrolizumab-advanced-renal-cell-carcinoma
7. Hahn AW, Klaassen Z, Agarwal N *et al.* First-line treatment of metastatic renal cell carcinoma: a systematic review and network meta-analysis. *Eur. Urol. Oncol.* 2(6), 708–715 (2019).
8. Riaz IB, He H, Ryu AJ *et al.* A Living, Interactive systematic review and network meta-analysis of first-line treatment of metastatic renal cell carcinoma. *Eur. Urol.* 80(6), 712–723 (2021).
9. Lombardi P, Filetti M, Falcone R *et al.* New first-line immunotherapy-based combinations for metastatic renal cell carcinoma: a systematic review and network meta-analysis. *Cancer Treat. Rev.* 106, DOI: 10.1016/j.ctrv.2022.102377 (2022).
10. Qu H, Mu Z, Wang K, Hu B. A network meta-analysis of the differences in effectiveness and safety between nivolumab and targeted drug therapy in metastatic renal cell carcinoma. *J. Oncol.* 2022, DOI: 10.1155/2022/5805289 (2022).
11. Bhatnagar N, Lakshmi PV, Jeyashree K. Multiple treatment and indirect treatment comparisons: an overview of network meta-analysis. *Perspect. Clin. Res.* 5(4), 154–158 (2014).
12. Monnickendam G, Zhu M, Mckendrick J, Su Y. Measuring survival benefit in health technology assessment in the presence of nonproportional hazards. *Value Health* 22(4), 431–438 (2019).
13. Freeman SC, Cooper NJ, Sutton AJ, Crowther MJ, Carpenter JR, Hawkins N. Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: application to a melanoma network. *Stat. Methods Med. Res.* 31(5), 839–861 (2022).
14. Freeman SC, Sutton AJ, Cooper NJ. Uptake of methodological advances for synthesis of continuous and time-to-event outcomes would maximize use of the evidence base. *J. Clin. Epidemiol.* 124, 94–105 (2020).
15. Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Med. Res. Methodol.* 11, 61 (2011).
16. Ouwens MJ, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Res. Synth. Methods* 1(3–4), 258–271 (2010).
17. NICE. Avelumab with axitinib for untreated advanced or metastatic renal cell carcinoma (2019). www.nice.org.uk/guidance/ta645/evidence
18. HAS. OPDIVO 10mg/ml (nivolumab) - carcinome à cellules rénales avancé, en association cabozantinib (2021). www.has-sante.fr/jcms/p_3297501/fr/opdivo-10-mg/ml-nivolumab-carcinome-a-cellules-renales-avance-en-association-cabozantinib
19. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl. Stat.* 43(3), 429–467 (1994).
20. Elaidi R, Phan L, Borchiellini D *et al.* Comparative efficacy of first-line immune-based combination therapies in metastatic renal cell carcinoma: a systematic review and network meta-analysis. *Cancers* 12(6), 1673 (2020).

21. Mihajlovic J, Postma MJ. Network meta-analysis of survival data using fractional polynomials - an example with first line metastatic renal cell cancer treatments. *Value Health* 18(7), A343 (2015).
22. McGregor BA, Petersohn S, Klijn S *et al.* Network meta-analysis (NMA) of first-line advanced renal cell carcinoma (1L aRCC) treatments: development of a decision algorithm for fractional polynomial (FP) model selection. *J. Clin. Oncol.* 40(Suppl. 6), 391–391 (2022).
23. Kraan CW, Nientker L, May J *et al.* PCN21 efficacy and safety in previously untreated, advanced/metastatic renal cell carcinoma – a systematic literature review update. *Value Health* 23, S424 (2020).
24. Heng DY, Xie W, Regan MM *et al.* Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor-targeted agents: results from a large, multicenter study. *J. Clin. Oncol.* 27(34), 5794–5799 (2009).
25. WebPlotDigitizer; Tohatgi, A. (2017). http://arohatgi.info/WebPlotDigitizer/index.html
26. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med. Res. Methodol.* 12, 9 (2012).
27. Motzer R, Alekseev B, Rha SY *et al.* Lenvatinib plus pembrolizumab or everolimus for advanced renal cell carcinoma. *N. Engl. J. Med.* 384(14), 1289–1300 (2021).
28. Rini BI, Plimack ER, Stus V *et al.* Pembrolizumab (pembro) plus axitinib (axi) versus sunitinib as first-line therapy for advanced clear cell renal cell carcinoma (ccRCC): results from 42-month follow-up of KEYNOTE-426. *J. Clin. Oncol.* 39(Suppl. 15), 4500–4500 (2021).
29. Choueiri TK, Motzer RJ, Rini BI *et al.* Updated efficacy results from the JAVELIN renal 101 trial: first-line avelumab plus axitinib versus sunitinib in patients with advanced renal cell carcinoma. *Ann. Oncol.* 31(8), 1030–1039 (2020).
30. Choueiri TK, Powles T, Burotto M *et al.* Nivolumab plus cabozantinib versus sunitinib for advanced renal-cell carcinoma. *N. Engl. J. Med.* 384(9), 829–841 (2021).
31. Hoaglin DC, Hawkins N, Jansen JP *et al.* Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. *Value Health* 14(4), 429–437 (2011).
32. R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria (2020).
33. Stan Development Team. RStan: the R interface to Stan (2021). http://mc-stan.org/users/interfaces/rstan
34. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J. Comput. Graphical Stat.* 7(4), 434–455 (1998).
35. Roy V. Convergence diagnostics for Markov chain Monte Carlo. *Ann. Rev. Stat. App.* 7, 387–412 (2020).
36. Motzer RJ, Robbins PB, Powles T *et al.* Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. *Nat. Med.* 26(11), 1733–1741 (2020).
37. Gerlinger M, Rowan AJ, Horswell S *et al.* Intratumor Heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366(10), 883–892 (2012).
38. Beksac AT, Paulucci DJ, Blum KA, Yadav SS, Sfakianos JP, Badani KK. Heterogeneity in renal cell carcinoma. *Urol. Oncol.* 35(8), 507–515 (2017).
39. Mcgregor BA, Geynisman DM, Burotto M *et al.* Efficacy outcomes of nivolumab + cabozantinib versus pembrolizumab + axitinib in patients with advanced renal cell carcinoma (aRCC): matching-adjusted indirect comparison (MAIC). *J. Clin. Oncol.* 39(Suppl. 15), 4578–4578 (2021).
40. Phillippo DM, Dias S, Ades AE *et al.* Multilevel network meta-regression for population-adjusted treatment comparisons. *J. R. Stat. Soc. Ser. A Stat. Soc.* 183(3), 1189–1210 (2020).
41. Guadalupi V, Carteni G, Iacovelli R *et al.* Second-line treatment in renal cell carcinoma: clinical experience and decision making. *Ther. Adv. Urol.* 13, DOI: 10.1177/17562872211022870 (2021).
42. Ouwens Mario. Novel approaches to common challenges in modelling of survival outcomes for cost-effectiveness analyses in oncology. *W8 Advanced Workshop, ISPOR 2018.* Barcelona, Spain (2018). www.ispor.org/docs/default-source/presentations/91714pdf.pdf?sfvrsn=a5b2756f_0
43. Wiecek W, Karcher H. Nivolumab versus cabozantinib: comparing overall survival in metastatic renal cell carcinoma. *PLoS One* 11(6), e0155389 (2016).
44. Toor K, Middleton MR, Chan K, Amadi A, Moshyk A, Kotapati S. Comparative efficacy and safety of adjuvant nivolumab versus other treatments in adults with resected melanoma: a systematic literature review and network meta-analysis. *BMC Cancer* 21(1), 3 (2021).
45. Cope S, Chan K, Campbell H *et al.* A Comparison of alternative network meta-analysis methods in the presence of nonproportional hazards: a case study in first-line advanced or metastatic renal cell carcinoma. *Value Health* 26(4), 465–476 (2023).