# Exploring hidden survival heterogeneity among first-line intermediate/poor-risk advanced renal cell carcinoma patients treated with nivolumab plus ipilimumab in the CheckMate 214 trial via parametric mixture models

MSR54

**Matthias Hunger,[1] Saby George,[2] Matthew Dyer,[3] Flavia Ejzykowicz,[4] Jessica R May,[3] Murat Kurt[4]**

1 ICON plc, Munich, Germany; 2 Roswell Park Comprehensive Cancer Center, Buffalo, NY, United States; 3 Bristol Myers Squibb, Uxbridge, United Kingdom; 4 Bristol Myers Squibb, Princeton, NJ, United States

## Background

- Renal cell carcinoma (RCC) is the most common type of kidney cancer in adults, accounting for 80-90% of all kidney malignancies. Historically, the five-year survival rate for those diagnosed with metastatic, or advanced RCC (aRCC) is 13%.[1]
- Approximately 77% of aRCC cases are classified as intermediate/poor-risk (I/P) according to International Metastatic Renal Cell Carcinoma Database Consortium (IMDC) model.[2,3] Nivolumab plus ipilimumab (NIVO+IPI) combination therapy has been approved by Food and Drug Administration in the US and the European Medicines Agency for the first-line (1L) treatment of aRCC patients with I/P-risk based on the results from pivotal, phase 3 randomized CheckMate 214 trial.[4]
- Analyses from the CheckMate 214 trial with a minimum 60 months of follow-up showed superior efficacy outcomes for NIVO+IPI compared with sunitinib in the I/P-risk population.[5] Specifically, the OS hazard ratio (HR) was 0.68 (95% confidence interval [CI], 0.58–0.81) and progression-free survival (PFS) (assessed per independent radiology review) HR was 0.73 (95% CI, 0.61–0.87).[5]
- Patterns of durable response and survival for patients treated with immuno-oncologic (IO) agents pose challenges for traditional survival modelling approaches. Models that do not account for survival heterogeneity may fall inadequate in capturing the changes in the complex hazard trend over time.[6-8]
- Survival heterogeneity and the underlying fraction of long-term survivors (LTS) among NIVO+IPI patients were previously explored by the applications of mixture cure models (MCMs). Analysis of PFS data estimated approximately 30% of the patients to be progression-free LTS.[9] On the other hand, analyses of OS and duration of response data for the responder subgroup showed that between 60% and 75% of the patients who achieve response are expected to be LTS.[10]
- Parametric Mixture Models (PMMs) are flexible survival analysis frameworks that can be used to explore clinically unobservable survival heterogeneity in a given population by modeling its survival as a mixture of the survival of two (or more) distinct and latent subgroups.[11] PMMs are increasingly considered as alternatives to standard parametric models for modeling long-term survival data for IO agents by allowing for varying degrees of flexibility to capture potential delays in response and heavy-tailed behavior of Kaplan-Meier (KM) curves.[8,12,13]

## Objective

- To visualize unobservable heterogeneity in survival among 1L I/P-risk aRCC patients treated with NIVO+IPI in the Checkmate 214 trial using PMMs
- To develop an algorithm that can guide the selection of PMMs for the evaluation of long-term PFS and OS by accounting for a variety of statistical metrics and clinical plausibility

## Methods

- The study population was assumed to consist of two non-overlapping and exhaustive latent subgroups with distinct survival patterns. Between the two subgroups, the one with more favorable estimated survival is referred to as “high-performers”, and the other is referred to as the “low-performers”.
- In its simplest form, the survival function of the population using PMMs [$S(t)$] can be structurally expressed as the weighted average of the survival of high- and low-performers (subgroups 1 and 2 below, respectively, without loss of generality):

  $$S(t) = p * S_1(t) + (1-p) * S_2(t)$$, where

  $p$ and $(1\text{-}p)$ represent the estimated proportions of patients categorized as high- and low-performers, respectively. Estimated survival functions of high- and low-performers are denoted by $S_1(t)$ and $S_2(t)$, respectively, where $S_1(t) \geq S_2(t)$ for all t.
- PMMs were fitted separately to PFS and OS data from the trial with minimum 60-months of follow-up to simultaneously elicit the proportion and survival function of each subgroup.
- For each subgroup’s survival, candidate parametric distributions recommended by the National Institute for Health and Care Excellence (NICE) (Exponential, Weibull, Log-Logistic, Log-Normal, and Gamma) were considered. Gompertz and Generalized Gamma distributions were omitted from consideration to avoid the risk of over-fitting. In total, 15 PMMs were fitted to each endpoint using the *fmm* package in Stata 14.
- General population mortality rates based on United Kingdom (UK) Office of National Statistics life tables (matched to the trial population by baseline age and gender) were used to gauge the clinical plausibility of the estimated survival in the high-performers subgroup.
- The potential for a local versus global solution in the likelihood maximization was evaluated by testing extreme starting values for the expectation-maximization algorithm employed by the *fmm* package in Stata.
- Best-fitting PMMs among 15 different combinations were identified based on their statistical goodness-of-fit measures and visual inspection using a two-step algorithm:
  - In Step 1, all PMMs dissatisfying at least one of the following conditions were eliminated: **A)** Survival of the high-performer subgroup was no more than 2% higher than general population OS at any point in time, **B)** Maximum deviation between the estimated and observed KM curves for the overall study population was < 5% at all times, **C)** Estimated survival of the subgroups do not cross each other at any point in time. Only the PMMs satisfying criteria A), B), and C) together were considered for additional filtering in Step 2.
  - In Step 2, the applied filtering criteria were more qualitative and included comparisons based on Akaike/Bayesian Information Criterion (AIC/BIC), estimated sizes of subgroups (mixtures with drastically imbalanced subgroup weights were penalized), shapes of long-term extrapolations for subgroups, discrepancies between estimated and observed hazards for the overall study population, and consistency of results with the extreme value testing.

## Results

**Analysis of PFS:**

- Among all 15 candidate PMMs, eight combinations were deemed viable, by satisfying all model selection criteria (Table 1). Estimated survival curves from these models (for the overall study population and the two latent subgroups) along with the reported KM-curve are shown in Figure 1.

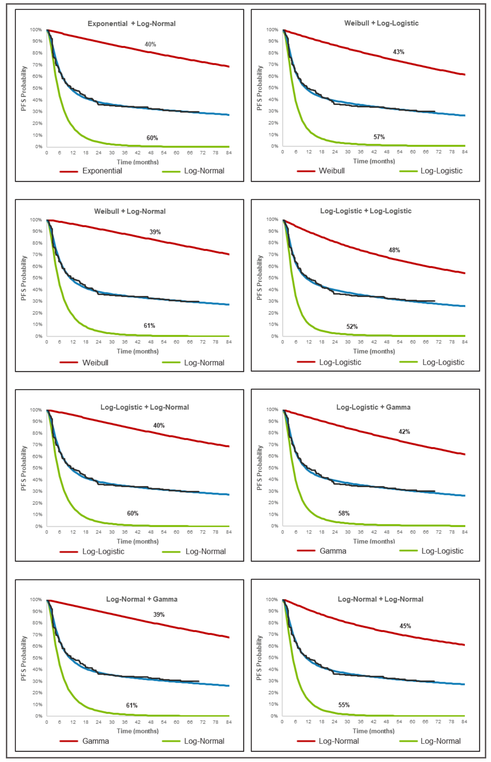
Table 1. Results of model-selection algorithm for PFS

| Model | Exponential + Exponential | Exponential + Weibull | Exponential + Log-Logistic | Exponential + Log-Normal | Exponential + Gamma | Weibull + Weibull | Weibull + Log-Logistic | Weibull + Log-Normal | Weibull + Gamma | Log-Logistic + Log-Logistic | Log-Logistic + Log-Normal | Log-Logistic + Gamma | Log-Normal + Log-Normal | Log-Normal + Gamma | Gamma + Gamma |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High-performer survival vs. general population OS | ✓ | X | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Estimated vs. observed KM-curves for the overall study population | X | X | X | ✓ | X | X | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | X |
| Crossing of survival functions of high- and low-performers | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| AIC/BIC | | | | | | | | | | + | | | +++ | ++ | |
| Estimated subgroup sizes | | | | | | | | | | | | | | | |
| Shapes of long-term extrapolations for both subgroups | | | | | | | | | | | | | | | |
| Estimated vs. observed hazards for the overall study population | | | | | | | | | | | | | | | |
| Consistency of results with extreme-value testing | | | | - | | | - | - | | | - | | | | |

In part 1 (upper part of the table), distributions fulfilling/not fulfilling a criterion are shown by a green check mark/a red cross. In part 2 (lower part of the table), remaining distributions are shown in light green, and +/- symbols are used to indicate the quality of the fits relative to each other with more +/- symbols showing better/worse fits.

- Majority of the viable combinations included a Log-Normal model (5 out of 8); best overall fit to the PFS data was provided by a mixture of two Log-Normal distributions (weights: 45% for high-performer subgroup, 55% for low-performer subgroup). Second- and third-best fits were obtained by a Log-normal + Gamma mixture (weights: 39% for high-performer subgroup, 61% for low-performer subgroup) and a Log-Logistic + Log-Logistic mixture (weights: 48% for high-performer subgroup, 52% for low-performer subgroup), respectively.
- Across all viable PMMs, the estimated weight of the high-performer subgroup ranged between 39% and 48% showing stability of PMMs in capturing the heterogenous PFS behavior between the subgroups.

Figure 1. Best-fitting PFS curves for the overall study population and for each latent subgroup from the PMMs




The red (green) line depicts the elicited PFS function in the high (low)-performers subgroup. The blue line depicts estimated PFS in the overall population (weighted average of PFS in the two subgroups). Observed PFS is shown by the KM curve in black. Percentages shown along the estimated survival functions of subgroups refer to their corresponding weights.

- Estimated 5-year restricted mean PFS across the selected top 3 PMMs showed negligible differences (<0.6 month), ranging from 25.8 to 26.4 months.

**Analysis of OS:**

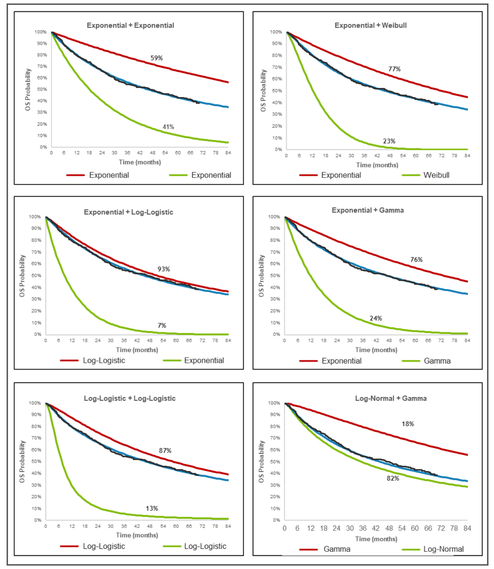
- Among all 15 candidate PMMs, six combinations were deemed viable, by satisfying all model selection criteria (Table 2). Estimated survival curves from these models (for the overall study population and the two latent subgroups) are shown in Figure 2.
- Majority of the viable combinations included an Exponential model (4 out of 6). Consistent with this prevalence, best overall fit to the OS data was provided by a mixture of two Exponential distributions (weights: 59% for high-performer subgroup, 41% for low-performer subgroup).
- Second and third best fits were obtained by an Exponential + Weibull (weights: 77% for high-performer subgroup, 23% for low-performer subgroup) and an Exponential + Log-Logistic mixture (weights: 93% for high-performer subgroup, 7% for low-performer subgroup), respectively.
- Across all viable PMMs, the estimated weight of the high-performer subgroup ranged between 18% and 93%. Compared to PFS, PMMs were less stable in exploring heterogeneous OS behavior between the subgroups.

Table 2. Results of model-selection algorithm for OS

| Model | Exponential + Exponential | Exponential + Weibull | Exponential + Log-Logistic | Exponential + Log-Normal | Exponential + Gamma | Weibull + Weibull | Weibull + Log-Logistic | Weibull + Log-Normal | Weibull + Gamma | Log-Logistic + Log-Logistic | Log-Logistic + Log-Normal | Log-Logistic + Gamma | Log-Normal + Log-Normal | Log-Normal + Gamma | Gamma + Gamma |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High-performer survival vs. general population OS | ✓ | ✓ | ✓ | | ✓ | X | X | | ✓ | ✓ | | ✓ | | ✓ | ✓ |
| Estimated vs. observed KM-curves for the overall study population | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | X | ✓ | | X | | ✓ | X |
| Crossing of survival functions of high- and low-performers | ✓ | ✓ | ✓ | | ✓ | ✓ | X | | ✓ | ✓ | | ✓ | | ✓ | ✓ |
| AIC/BIC | + | | | | | | | | | | | | | | |
| Estimated subgroup sizes | | | - - - | | | | | | | - - | | | | | |
| Shapes of long-term extrapolations for both subgroups | | | | | | | | | | | | | | | |
| Estimated vs. observed hazards for the overall study population | | | | | | | | | | | | | | | |
| Consistency of results with extreme-value testing | | - | - | | - | | | | | - | | | - | | |

In part 1 (upper part of the table), distributions fulfilling/not fulfilling a criterion are shown by a green check mark/a red cross. In part 2 (lower part of the table), remaining distributions are shown in light green, and +/- symbols are used to indicate the quality of the fits relative to each other with more +/- symbols showing better/worse fits.
The 4 PMMs with an entire white-colored columns were eliminated from the beginning due to violation of the strict condition on the estimated sizes (to be at least 5%) of subgroups.

Figure 2. Best-fitting OS curves for the overall study population and for each latent subgroup from the PMMs




The red (green) line depicts the elicited OS function in the high (low)-performers subgroup. The blue line depicts estimated OS in the overall population (weighted average of OS from the two subgroups). Observed OS is shown by the KM curve in black. Percentages shown along the estimated survival functions of subgroups refer to their corresponding weights.

- Estimated 5-year restricted mean OS across the selected top 3 PMMs showed negligible differences (< 0.8 month), ranging from 38.7 to 39.5 months.
- Across the corresponding top 3 PMMs, long-term (30-year) background mortality-adjusted mean PFS and OS ranged between 6 and 6.68 years, and between 7.49 and 8.12 years, respectively.

## Discussion

- CheckMate 214 has the longest follow-up data across all randomized phase 3 trials investigating 1L treatment of aRCC patients with IO agents. Sustained plateau behavior in the KM-curves for PFS, high response rates and durability of response reported from the 5-year follow-up for patients treated with NIVO+IPI are strong clinical indications of survival heterogeneity.
- The distinct nature of the survival curves estimated for each latent class and the differences in their weights are indicative of strong statistical ability of PMMs in eliciting survival heterogeneity that may not be clearly manifested by the prognostic variables for the trial population.
- Unlike the strong plateau behavior in the KM-curve for PFS, there was no apparent flattening in the tail of the KM-curve for OS at 60 months of follow-up. Therefore, the range of fractions of high-performer subgroup was wider in the estimations from the OS data than those from the PFS data.
- Estimated 5-year restricted mean PFS and OS were consistent across the top-3 fitting PMMs, highlighting the robustness of models satisfying the filtering criteria.
- The heterogeneity in the PFS outcomes explored by the PMMs were consistent with those previously reported from the MCMs, a special but more restrictive group of mixture models where the cured subgroup’s survival was assumed to be driven only by general population mortality rates and need not be estimated by the model [9].
- While MCMs are built with the notion of cure, a limitation of the PMMs is the lack of clinical identification behind latent subgroups. Therefore, it may not be possible to validate the estimated survival functions of the subgroups from a clinical perspective or against external data.
- This analysis was limited to I/P-risk patients treated with NIVO+IPI in the CheckMate 214 trial due to novel mechanism of action for IO agents. For the sunitinib arm, exploration of PMMs were not considered as there is no expectation of heterogeneity in survival or durability of response based on treatment’s mechanism of action or prognostic clinical variables.
- With the exception of the Exponential + Exponential mixture model, all PMMs considered in this analysis require > 3 parameters whereas the majority of the standard parametric models require < 3 parameters. Therefore, the flexibility to tackle heterogeneous survival data by PMMs comes with a trade-off of over-fitting of the data which may lead to overly-optimistic long-term survival projections.

## Conclusions

- PMMs may adequately capture the complex survival and hazard trends for 1L I/P-risk aRCC patients treated with NIVO+IPI by offering insights on potential survival heterogeneity without making clinically restrictive assumptions.
- From a statistical standpoint, as observed for 4 and 5 candidate combinations in the analyses of PFS and OS data, respectively, optimization of the inherently complex likelihood functions of PMMs may be computationally intensive due to potential convergence issues.
- Capturing the time-varying hazard trend in the OS data with an Exponential + Exponential mixture model can relax the need for tunnel states in OS predictions for I/P-risk RCC populations in earlier treatment settings who are treated with NIVO+IPI upon recurrence.

## References

1. Surveillance, Epidemiology, and End Results Program, 2010-2016 database. https://seer.cancer.gov/statfacts/html/kidrp.html
2. Heng, D.Y. et al. *J Clin Oncol* 2009; 27(34):5794-5799
3. Heng, D.Y. et al. *Lancet Oncol* 2013; 14(2):141-148.
4. Motzer, R.J. et al. *N Engl J Med* 2018; 378(14):1277-1290.
5. Motzer, R.J. et al. *Ann Oncol* 2021; 32(suppl 5):S685-S687.
6. Rutherford M.J. et al. NICE DSU Technical Support Document 21 https://www.sheffield.ac.uk/media/34188/download?attachment
7. Latimer N.R. *Med Decis Making* 2013; 33(6):743-754.
8. Paly, VF. *MDM Policy & Pract* 2022; 7(1):1-15
9. Paly, V.F. et al. *Value Health* 2022; 25(suppl 1):S210.
10. George, S. et al. *Ann Oncol* 2022; 33(suppl 7):S1210-S1211
11. Klijn, S. et al. *Pharmacoeconomics* 2021; 39(3):345-356
12. Ouwens, M. et al. *Pharmacoeconomics* 2019; 37(9):1129-1138
13. Cislo et al. *Value Health* 2021; 24(11):1643-1650
14. StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC

## Acknowledgments

All authors contributed to and approved the presentation. Analytical and writing assistance was provided by ICON, funded by Bristol Myers Squibb.

Presented at ISPOR 2023; May 7-10, 2023; Boston, MA

Email: murat.kurt@bms.com