PCN440

# Network Meta-Analysis Using Fractional Polynomials: Heuristic for Model Selection Incorporating Beyond-Trial Extrapolations (A Melanoma Example)

Sarah Goring,[1] Kabirraaj Toor,[1] Dieter Ayers,[1] Keith Chan,[1] Shannon Cope,[1] Helen M. Johnson,[2] Andriy Moshyk,[3] Murat Kurt,[3] Srividya Kotapati,[3] Jeroen P. Jansen[4]

[1]Precision Xtract, Vancouver, BC, Canada; [2]Bristol-Myers Squibb, Uxbridge, UK; [3]Bristol-Myers Squibb, Princeton, NJ, USA; [4]Precision Xtract, Oakland, CA, USA

## Background

- Fractional polynomials (FP) provide flexible parametric modeling approaches for analyzing time-to-event data
- In FP models, the log hazard of an event, such as all-cause mortality, is fit as a function of time using parametric models of the form

$$\text{log hazard} = (\mu_0) + (\mu_1)\, t^{P1} + (\mu_2)\, t^{P2},$$

where $\mu_0$ represents the scale parameter, $\mu_1$ and $\mu_2$ represent shape parameters, $t$ represents time, and $P1$ and $P2$ are powers from the set (eg, {−1, −0.5, 0, 0.5, 1}), and where $t^0 = \ln(t)$. When P1 = P2 = P, the model becomes a repeated powers model

$$\text{log hazard} = (\mu_0) + (\mu_1)\, t^{P} + (\mu_2)\, t^{P} \ln(t)$$

- To capture treatment effects, differences (d) can be added to the first, second, and/or third term of FP models

$$\text{log hazard} = (\mu_0 + d_0) + (\mu_1 + d_1)\, t^{P1} + (\mu_2 + d_2)\, t^{P2}$$

  - These differences can be incorporated into meta-analyses or network meta-analyses (NMAs); for treatment $i$, pooled estimates for $d_{0,i}$, $d_{1,i}$, and/or $d_{2,i}$ can be produced relative to a reference treatment[1]
- These models produce time-dependent hazard ratios (HRs) and therefore are used in networks for which the proportional hazards assumption does not hold
- Despite the benefits of using FP models, there are some challenges in the modeling process
  - Fitting data to all possible model forms produces a large number of candidate models
  - Functional forms that fit data well in the early course of follow-up may have clinically implausible behavior in the tails (when evidence is based on a small at-risk population with great uncertainty) and during extrapolated periods
    - In networks of evidence with heterogeneous durations of follow-up, care must be taken in distinguishing between time periods for which HRs were observed and extrapolated; the clinical plausibility of the model forms should be evaluated
  - The deviance information criterion (DIC), despite its widespread use, suffers from a lack of consistency and an arguably poor theoretical justification[2]
    - Therefore, it is unclear whether it sufficiently penalizes FP model complexity; it can thus lead to overfitting, particularly in the tails. Furthermore, it cannot be applied to extrapolated periods

## Objective

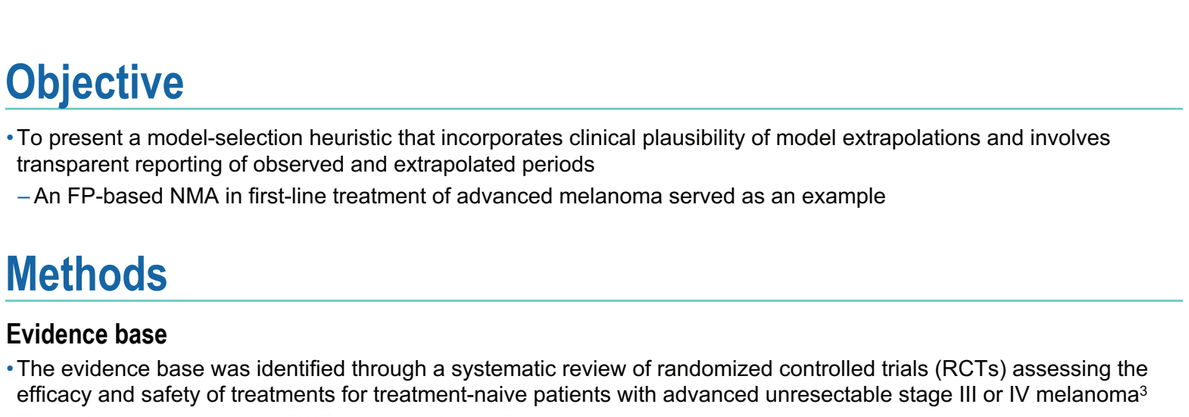
- To present a model-selection heuristic that incorporates clinical plausibility of model extrapolations and involves transparent reporting of observed and extrapolated periods
  - An FP-based NMA in first-line treatment of advanced melanoma served as an example

## Methods

### Evidence base

- The evidence base was identified through a systematic review of randomized controlled trials (RCTs) assessing the efficacy and safety of treatments for treatment-naive patients with advanced unresectable stage III or IV melanoma[3]
- The network formed by the RCTs is presented in **Figure 1**
- A time-varying HR analysis was conducted, as the proportional hazard assumption did not appear to hold
- We distinguished between observed and extrapolated periods for each pair of comparators in the network
  - First, for each edge in the network, the maximum directly observed period was established as the maximum duration of follow-up across all RCTs in that particular connection (denoted below each RCT name)
  - Second, the shortest path (the path connecting the nodes with the fewest possible edges) between each pair of comparators was established
  - Third, as a common duration of follow-up, the minimum of the directly observed periods along the shortest network path was calculated; subsequent time points were considered as extrapolations
    - In cases in which ≥ 2 paths were shortest (eg, route 1 and route 2 in **Figure 1**), the route with the longest common duration of follow-up was taken
- As an example, the extrapolated period for nivolumab plus ipilimumab (NIVO+IPI) versus dabrafenib plus trametinib (DAB+TRAM) begins after 46 months, which is the common duration of follow-up time along route 2 in **Figure 1**

**Figure 1. Network of evidence for treatment-naive advanced unresectable melanoma[a]**




[a]For ENCO, IPI, and NIVO, numbers represent doses (in mg/kg).
[b]Reference treatment. Labels on connections indicate RCT name (maximum duration of follow-up).
BINI, binimetinib; COBI, cobimetinib; DAB, dabrafenib; DTIC, dacarbazine; ENCO, encorafenib; PEM, pembrolizumab; TRAM, trametinib; VEM, vemurafenib.

### Fractional polynomial models

- Using digitally reconstructed individual patient-level data for overall survival, we fit Weibull-based FP models (where shape 1, power P1 = 0) and Gompertz-based FP NMA models (where shape 1, power P1 = 1)[4]
- Treatment effects were placed on the scale parameter ($d_0$), representing proportional hazards (complexity level 1; **Figure 2**)
- To each model, we added further complexity (**Figure 2**)
  - Level 2: treatment effects were added to the shape 1 (ie, time-related) parameter ($d_1$)
  - Level 3: a second shape parameter was added (shape 2; P2 powers chosen from {−1, −0.5, 0, 0.5, 1})
  - Levels 4a and 4b: both shape parameters were retained and treatment effects were added to either shape 1 ($d_1$) or shape 2 ($d_2$)
  - Level 5: treatment effects were added to both shape parameters, $d_1$ and $d_2$
- Model families were defined by the choice of P1 or P2
- As there was typically only 1 RCT connecting any 2 given nodes in the network (**Figure 1**), fixed-effect models were selected a priori

**Figure 2. Fractional polynomial model development and candidate models**

| Complexity | Scale | Shape 1 (P1) | Shape 2 (P2) | Number of combinations[a] |
|---|---|---|---|---|
| 1 | $(\mu_0 + d_0)$ | $\mu_1 \ln(t)$ | None | 1 |
| | $(\mu_0 + d_0)$ | $\mu_1 t$ | None | 1 |
| 2 | $(\mu_0 + d_0)$ | $(\mu_1 + d_1) \ln(t)$ | None | 1 |
| | $(\mu_0 + d_0)$ | $(\mu_1 + d_1) t$ | None | 1 |
| 3 | $(\mu_0 + d_0)$ | $\mu_1 \ln(t)$ | $\mu_2 t^{P2a}$ | 5 |
| | $(\mu_0 + d_0)$ | $\mu_1 t$ | $\mu_2 t^{P2a}$ | 5 |
| 4a | $(\mu_0 + d_0)$ | $(\mu_1 + d_1) \ln(t)$ | $\mu_2 t^{P2a}$ | 5 |
| | $(\mu_0 + d_0)$ | $(\mu_1 + d_1) t$ | $\mu_2 t^{P2a}$ | 5 |
| 4b | $(\mu_0 + d_0)$ | $\mu_1 \ln(t)$ | $(\mu_2 + d_2) t^{P2a}$ | 5 |
| | $(\mu_0 + d_0)$ | $\mu_1 t$ | $(\mu_2 + d_2) t^{P2a}$ | 5 |
| 5 | $(\mu_0 + d_0)$ | $(\mu_1 + d_1) \ln(t)$ | $(\mu_2 + d_2) t^{P2a}$ | 5 |
| | $(\mu_0 + d_0)$ | $(\mu_1 + d_1) t$ | $(\mu_2 + d_2) t^{P2a}$ | 5 |

[a]The number of combinations is based on setting P2 to each value within the set {−1,−0.5,0,0.5,1}.

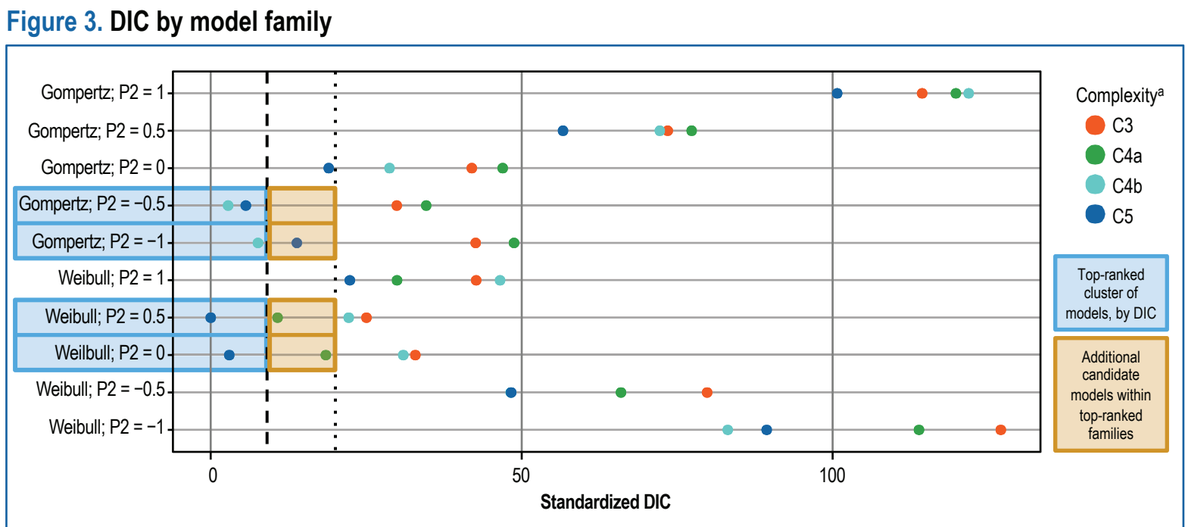
### Model-selection heuristic

- The model-selection heuristic involved ranking models according to 4 criteria
  1. **DIC:** The DIC for each model was standardized by subtracting the smallest DIC among all models. (Models in the top-ranked cluster [in this example, the top 5 models] according to DIC were retained as candidate models)
  2. **Additional model complexity penalties:** As the DIC may insufficiently penalize for complexity, we also retained less complex models from the same families identified by DIC alone. (To limit the number of candidate models, we excluded models with a standardized DIC > 20)
  3. **Visual inspection of observed versus modeled output:** Study-specific FP model survival curves were overlaid upon observed Kaplan–Meier (KM) curves to visualize deviations between observed and parameterized curves. (For transparency, extrapolations beyond the indirectly observed period were distinguished from observed periods using dashed lines)
  4. **Clinical plausibility of extrapolation periods:** For all time periods beyond the observed period, the clinical plausibility was evaluated based on observed data from other drugs within the same class and indication

## Results

- In total, 44 fixed-effect models were run
- According to the DIC, the 5 top-ranked models included 2 Weibull-based FP model families (shape 2, powers P2 = 0.5 and P2 = 0) and 2 Gompertz-based FP models (shape 2, powers P2 = −1 and P2 = −0.5) (**Figure 3**; blue shading)
  - Models with complexity levels 1 and 2 had standardized DICs > 350 (results not shown)
- Additional models from the same family as these 5 top-ranked models, whose standardized DICs were < 20, were also retained as candidate models (**Figure 3**; gold shading)

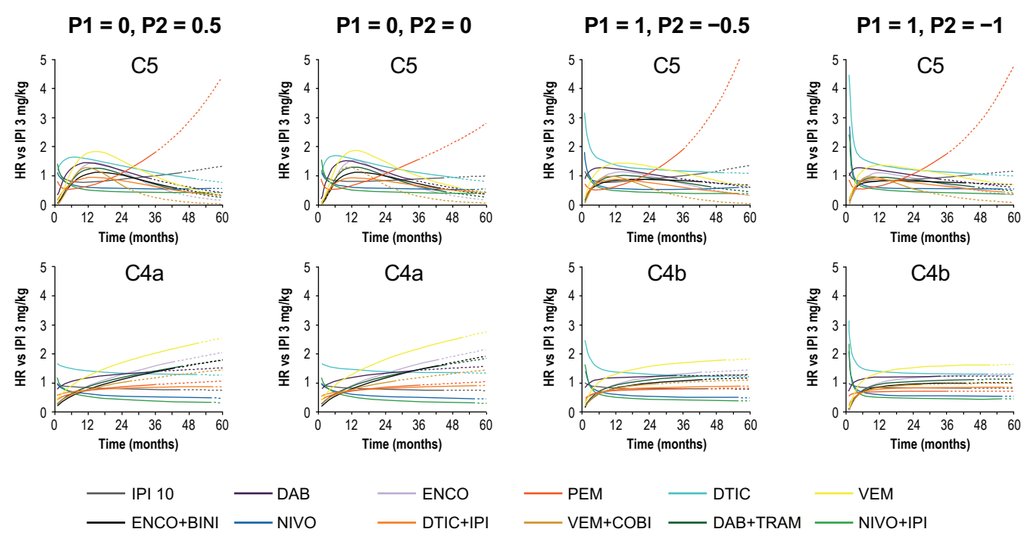
**Figure 3. DIC by model family**




[a]C3 involves a scale and 2 shape parameters, with treatment effect on the scale only; C4a involves a scale and 2 shape parameters, with treatment effect on the scale and the first shape parameter; C4b involves a scale and 2 shape parameters, with treatment effect on the scale and the second shape parameter; C5 involves a scale and 2 shape parameters, with treatment effect on the scale and both shape parameters (see **Figure 2**).

- For each candidate model, plots of the estimated HRs relative to the reference treatment (IPI 3) were used to distinguish between the observed and extrapolated periods (**Figure 4**)
- Visual inspection of these HRs revealed that the most complex models (C5) showed clinically implausible within-drug class heterogeneity in the extrapolated periods when compared with simpler models (C4a or C4b)
- Specifically regarding trends in extrapolated HRs versus IPI 3
  - For PEM, **extrapolated** HRs differed markedly from the **observed** HRs for NIVO
  - For VEM+COBI, **extrapolated** HRs differed from the **observed** HRs for other BRAF+MEK inhibitor combinations

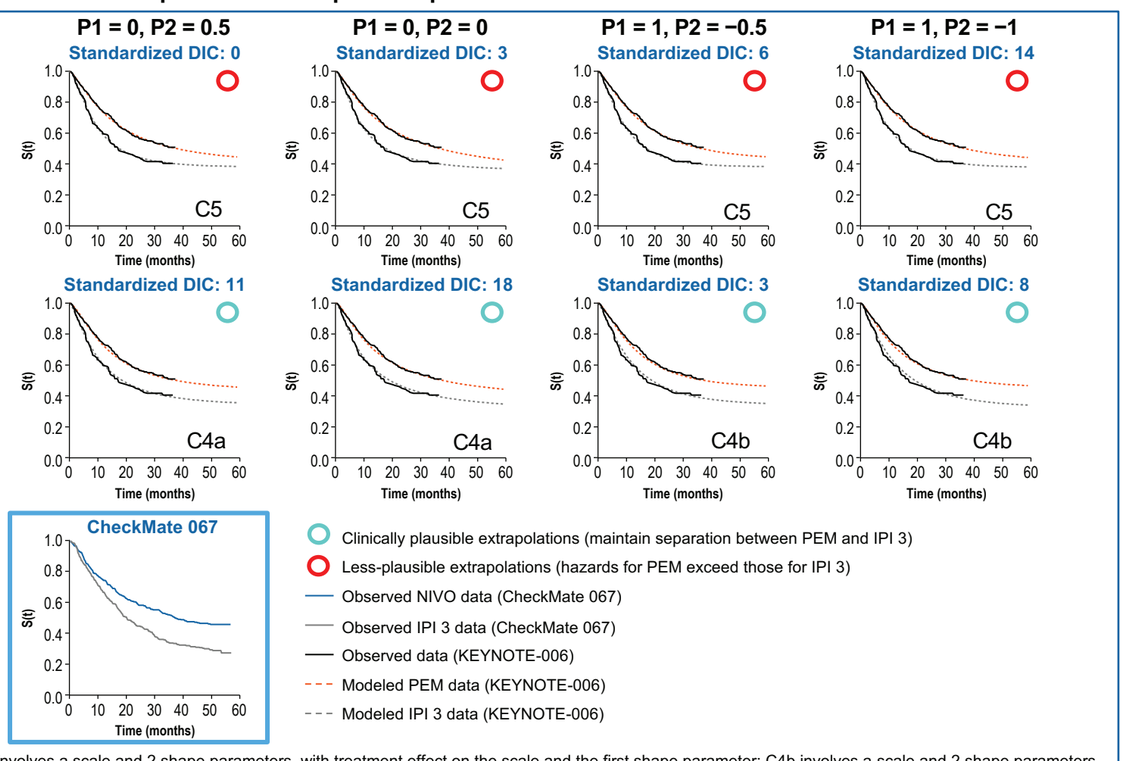
**Figure 4. Evaluation of clinical plausibility of HRs versus IPI 3 mg/kg for candidate models[a]**




[a]C4a involves a scale and 2 shape parameters, with treatment effect on the scale and the first shape parameter; C4b involves a scale and 2 shape parameters, with treatment effect on the scale and the second shape parameter; C5 involves a scale and 2 shape parameters, with treatment effect on the scale and both shape parameters.

- When reviewing the study-specific FP models, compared with observed data, differences were noted in how each FP model fit the tails and extrapolation periods (eg, the extrapolation period for KEYNOTE-006 is displayed in **Figure 5**)
  - In the complexity-level C5 models, the KM tails and extrapolation periods for PEM versus IPI 3 were converging (corresponding to an HR > 1 for PEM vs IPI 3), whereas the level C4a and C4b models maintained separation (corresponding to an HR of ~ 1)
  - The C5 models were more heavily influenced by events near the end of KEYNOTE-006 RCT follow-up, when the population at risk was small and mostly censored. (By 36 months, the numbers at risk were 12 patients and 4 patients for the PEM and IPI 3 arms, respectively)
  - These trends evident in the C5 models were inconsistent with observed trends in the same period for NIVO versus IPI 3
- The heuristic guided the final model choice toward less complex models with similar DICs, but without the C5 level of complexity

**Figure 5. Visual inspection of extrapolation periods for KEYNOTE-006[a]**




[a]C4a involves a scale and 2 shape parameters, with treatment effect on the scale and the first shape parameter; C4b involves a scale and 2 shape parameters, with treatment effect on the scale and the second shape parameter; C5 involves a scale and 2 shape parameters, with treatment effect on the scale and both shape parameters.

## Discussion

- Our heuristic is an integrated model-selection framework combining both internal validity using the statistical measures of a model fit to the observed data and external validity of the tail and extrapolation periods
  - This is consistent with good-practice guidance for survival extrapolations in cost-effectiveness models[5]
- Our findings highlight the caveat that the most complex (C5) models may overfit to the data at the end of the follow-up period when the number of patients at risk is typically low; such overfitting may yield clinically implausible extrapolations
  - Selecting models of slightly lower complexity, with 2 shape parameters but treatment effects on only 1 of these parameters (eg, C4a, C4b), appeared to be a more prudent approach
  - Our heuristic was based on an NMA involving only 36 months of follow-up data from KEYNOTE-006. Recently, 5-year results of KEYNOTE-006 were released, and the newly observed data were consistent with the clinically plausible C4a and C4b extrapolations; the final FP model in the updated NMA was C4b (P1 = 1; P2 = −1)[3,6]
- Clear demarcation of observed and extrapolated periods provided greater transparency in reporting, ensuring that the latter are not misrepresented, particularly in networks with heterogeneous follow-up times
  - The clarification that the extrapolated period begins at the shortest observed period **along the network path** is important, as this may be shorter than the observed period for 2 individual comparators
- Our heuristic was not prescriptive. Further research across multiple endpoints and indications is needed to ensure a specific predefined approach to model selection can be followed to minimize subjectivity in the model-selection process

## Conclusions

- FP models with 2 shape parameters and treatment effects on both (C5) may overfit the tails of the follow-up period, when numbers at risk tend to be small; less complex models (eg, C4a, C4b) can provide comparable model fit with greater clinical plausibility in the tails and extrapolation periods
- A combination of DIC, visual inspections, and consideration of external validity should be used as part of the model-selection heuristic
- When reporting time-specific HRs, a clear demarcation of observed and extrapolated periods should be used

## References

1. Jansen JP. *BMC Med Res Methodol* 2011;11:61.
2. Spiegelhalter DJ, et al. *J R Stat Soc Series B Stat Methodol* 2014;76:485–493.
3. Mohr P, et al. Poster presentation at ESMO 2019; September 27–October 1, 2019; Barcelona, Spain. Poster 1347P.
4. Guyot P, et al. *BMC Med Res Methodol* 2012;12:9.
5. Latimer NR. *Med Decis Making* 2013;33:743–754.
6. Robert C, et al. *Lancet Oncol* 2019;20:1239–1251.

## Acknowledgments

- Bristol-Myers Squibb (Princeton, NJ) and ONO Pharmaceutical Company Ltd. (Osaka, Japan)
- This study was supported by Bristol-Myers Squibb
- All authors contributed to and approved the presentation; editorial assistance was provided by Jessica Franciosi, PhD, and Andrea Lockett of StemScientific, an Ashfield Company, funded by Bristol-Myers Squibb

Presented at ISPOR Europe 2019; November 2–6, 2019; Copenhagen, Denmark

Email: kabirraaj.toor@precisionxtract.com