PT5

Structural Uncertainty in Evidence Synthesis: A Case Study Applying Model Averaging in Bayesian Multi-Level Network Meta-Regression

Sarah Goring[1], Dylan Maciel[2], Walter Bouwmeester[2], Shannon Cope[2], Jeroen P. Jansen[2,3], Harlan Campbell[2,4]

[1]SMG Outcomes Research, Vancouver, Canada, [2]Precision AQ; [3]University of California School of Pharmacy, San Francisco, USA; [4]University of British Columbia Dept of Statistics, Vancouver, Canada

Background

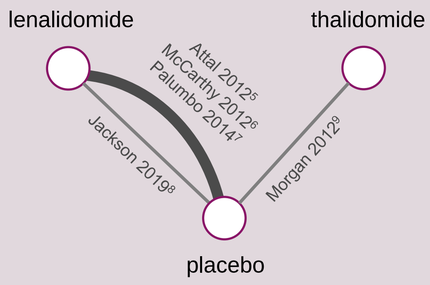
- Meta-analysis (MA) and its extension, network meta-analysis (NMA) provide a means for synthesizing all available evidence regarding relative treatment effects.
  - Multi-level network meta-regression (ML-NMR) models involve fitting parametric models to individual patient-level data (IPD) and pseudo-IPD, which incorporate covariates (prognostic factors and effect modifiers) for population adjustment.[1]
  - Structural uncertainty in terms of the choice of parametric model can be assessed with model averaging methods, such as Bayesian model averaging (BMA).[2]

Objective

- The aim was to explore model averaging techniques for a time-to-event analysis using ML-NMR based on a case study in newly diagnosed multiple myeloma (ndMM).

Methods

STEP 1: Network of evidence



Population: ndMM.

Intervention/Comparator: lenalidomide, thalidomide, and/or placebo.

Outcome: Progression-free survival (PFS).

Study design: RCTs* with IPD (▬) or aggregate-level pseudo-IPD (—) for PFS.

Abbreviations: RCT = randomized controlled trial.
** Identified by Leahy & Walsh 2019[3] & Phillippo et al. 2024,[4] representing constructed synthetic data; RCT references as cited by Leahy & Walsh 2019.[3]*

STEP 2: ML-NMR analysis

ML-NMR adjustment factors:

- Age
- International staging system stage
- Post-autologous stem cell transplant response
- Sex

Fit data to parametric forms:*

1. Lognormal
2. Log-logistic
3. Weibull (PH)
4. Weibull (AFT)
5. M-spline (7-knot)
6. Gompertz
7. Exponential
8. Gamma
9. Generalized gamma

Target population:

- Target population defined by patient characteristics in McCarthy 2012[6]

Outputs:

- Marginal difference (Δ) in RMST up to:
  - 4 years: maximum observed time†
  - 40 years: extrapolated time horizon

** Using multinma package in R (version 4.3.1)*
† Maximum observed time in the RCT having the shortest duration of follow-up (Attal 2012)[5]
Abbreviations: AFT = accelerated failure time; PH = proportional hazards; RMST = restricted mean survival time.

STEP 3: Model selection and averaging

Model selection was performed using three approaches:

1. "Best" model

- The 'best' model was selected using the leave-one-out information criterion (LOOIC),* similar to standard model selection in NMA.[4,10,11]

2. Averaging over models with "pseudo-BMA+" weights

- Averaging with pseudo-BMA+ weights is similar to standard BMA, but does not require calculating marginal likelihoods, which can be computationally expensive. Instead, the pseudo-BMA+ weights are based on the LOOIC of each model. As such, the 'best' model according to the LOOIC is given the highest weight. To reduce the risk of overfitting, the 'Bayesian bootstrap' regularizes the weights away from the extremes of 0 and 1.[12]

3. Stacking

- Stacking is a model averaging technique which seeks to optimize out-of-sample prediction and has been shown to outperform standard BMA in M-open settings (i.e., when the 'true' model is not amongst those in the list of candidate models).[12-13]

** Using the loo package in R (version 4.3.1)[14]*

Results

Figure 1 presents the LOOIC & model averaging weights for each model. **Figure 2** and **Figure 3** illustrate the predicted PFS & ΔRMST at 4 and 40 years in the target population for each approach.

1. "Best" model

- Of the 7 models that converged, log-logistic model was best fit to the data based on LOOIC.

2. Averaging over models with "pseudo-BMA+" weights

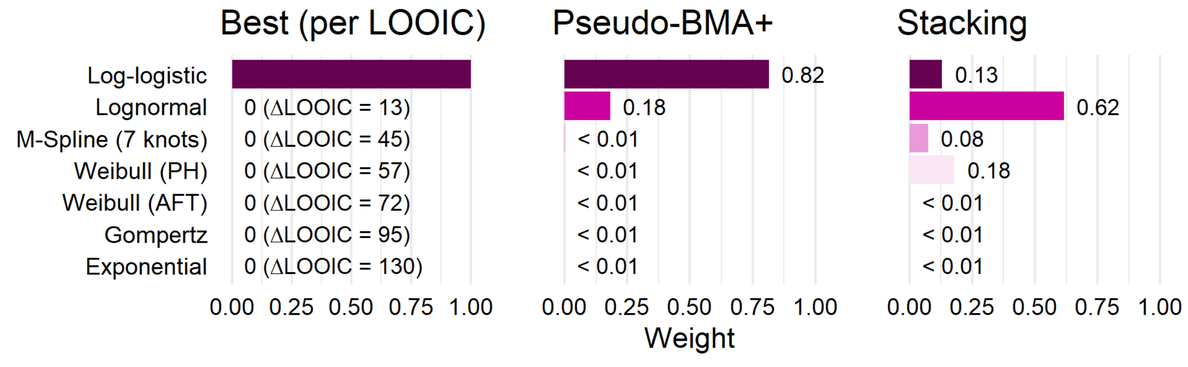
- Log-logistic model carried the largest weight (0.82), followed by lognormal (0.18); other models contributed negligible weight (<0.01).
- ΔRMST estimates were similar to best fitting (log-logistic) model estimates, although shifted toward the lognormal, with wider 95% credible intervals (CrIs).

3. Stacking

- Four models contributed to averaged estimates; lognormal model had largest weight (0.62).
- Generally, stacking estimates were consistent with the best model and pseudo-BMA+, with wider 95% CrIs.

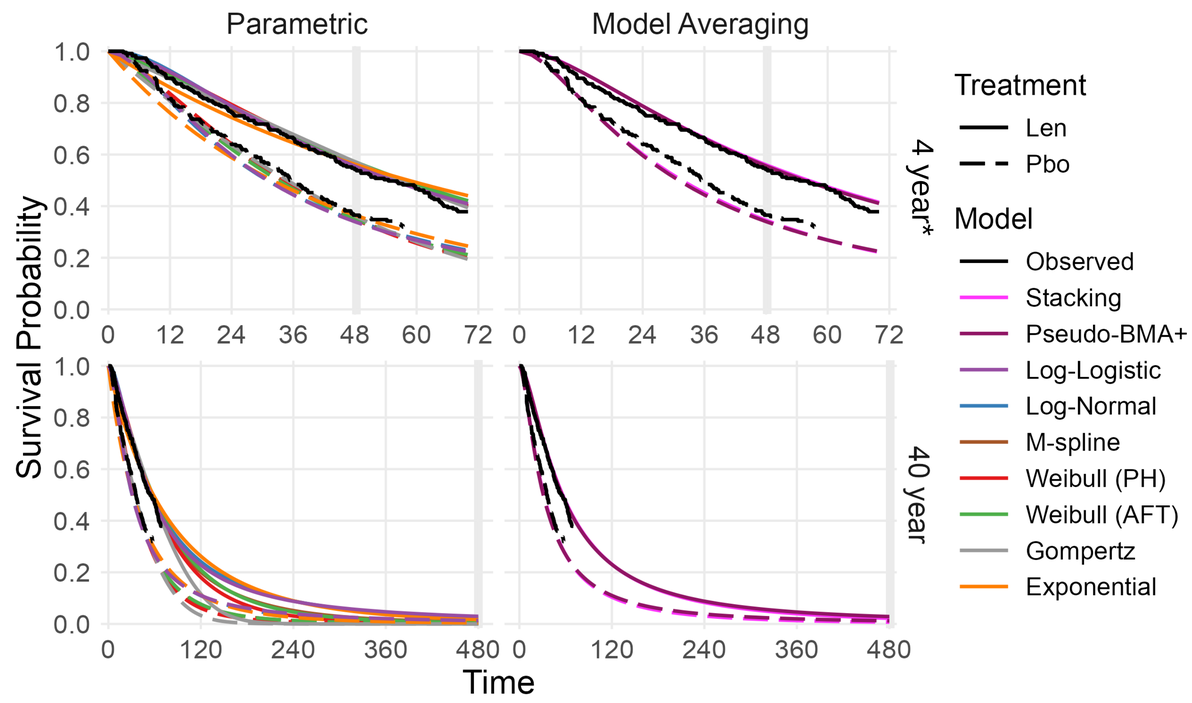
Results

Figure 1. Model averaging weights applied to each ML-NMR parametric form*



Abbreviations: AFT = accelerated failure time; BMA = Bayesian model averaging; ΔLOOIC = difference in leave-one-out information criterion relative to best; PH = proportional hazards.
** Gamma and generalized gamma models did not converge*

Figure 2. Kaplan-Meier PFS and marginal ML-NMR predictions in target population

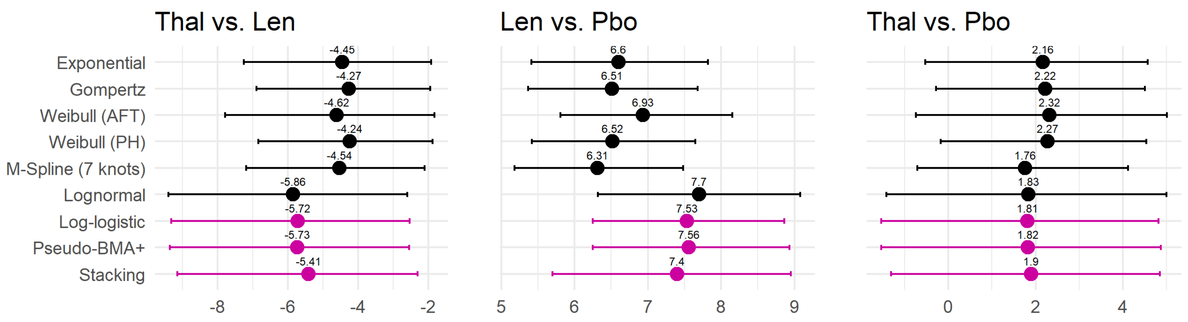


Abbreviations: AFT = accelerated failure time; BMA = Bayesian model averaging; Len = lenalidomide; ML-NMR = Multi-level network meta-regression; Pbo = placebo; PFS = progression-free survival; PH = proportional hazards; RCT = randomized controlled trial

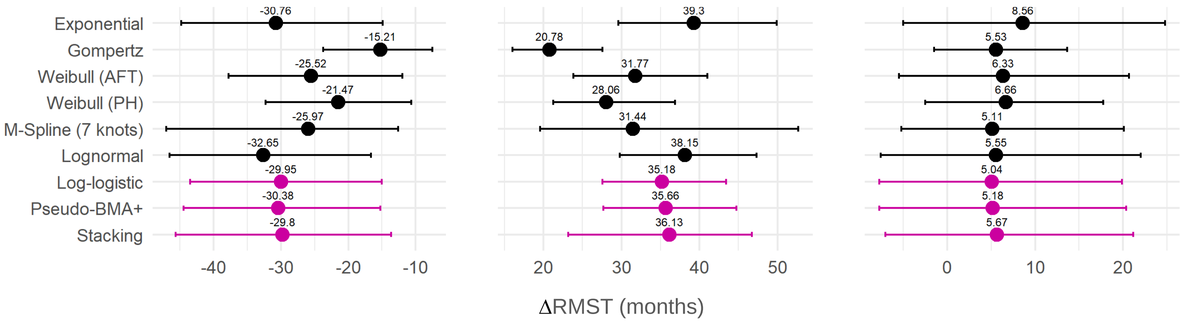
** 4 years represents the maximum observed period in Attal 2012[5] (the RCT having the shortest duration of follow-up); however, the observed period in McCarthy 2012[6] extends beyond this range (as shown in the observed data above).*

Figure 3. Model-specific, model-averaged & stacked marginal PFS ΔRMST in target population

4 years (maximum observed period*)



40 years (extrapolated period)



Abbreviations: AFT = accelerated failure time; BMA = Bayesian model averaging; Len = lenalidomide; PFS = progression-free survival; PH = proportional hazards; Pbo = placebo; RMST = restricted mean survival time; Thal = thalidomide.
** See footnote in Figure 2.*

Discussion

- **Results summary:** The pseudo-BMA+ and stacking methods appropriately capture structural uncertainty and resulted in wider credible intervals than with the 'best' model (per LOOIC).
  - Stacking approach is considered an optimal choice when the true model may not be captured within the set of candidate models,[12-13] which may be appropriate for ML-NMR case study.
- **Not explored:** Averaging across models with different covariates is possible, but covariate selection should be determined *a priori* based on literature review and clinical input.[15-17]
  - If exploring the role of covariates as a sensitivity, spike-and-slab priors may provide a single model with equivalent results to model averaging.[18]
- **Future research:** Case study model weights were informed by statistical fit to the observed period; however, information from clinical experts to narrow the subset of plausible models or inform model weights based on plausibility may be an area of future research.
- **Key message:** For ML-NMR, implementing pseudo-BMA+ and stacking is relatively straightforward and should be considered to address structural uncertainty for time-to-event outcomes.

References 1. Phillippo et al. 2020 *J Royal Statistical Society Series A: Statistics in Society*, *183*(3), 1189-1210; 2. Otten et al. 2023. *Pharmacoeconomics*. 41: 619–632; 3. Leahy & Walsh. 2019. *Res Synth Meth*. 10(4):546-568; 4. Phillippo et al. 2024. arXiv:240112640; 5. Attal et al. 2012. *N Engl J Med*. 366(19):1782-91; 6. McCarthy et al. 2012. *N Engl J Med*. 366(19):1770-81; 7. Palumbo et al. 2014. *N Engl J Med*. 371(10):895-905; 8. Jackson 2019 was cited as "Jackson GH, Davies FE, Pawlyn C, et al. Response adapted induction treatment improves outcomes for myeloma patients; results of the phase iii myeloma xi study; 2016"; 9. Morgan et al. 2012. *Blood*. 119(1):7-15; 10. Cope et al. 2022 *Value in Health*. 26(4):465-476; 11. Gallacher et al. 2021. *Med Decis Making* 41(4):476-484; 12. Yao et al. 2018. *Bayesian Anal*. 13(3): 917-1003; 13. Le & Clark. 2017. *Bayesian Anal*. 12(3): 807-829; 14. Vehtari & Gabry 2024 https://mc-stan.org/loo/articles/loo2-weights.html; 15. Freitag et al. 2023. *J Comp Eff Res*, *12*(10), e230046; 16. Phillippo et al. 2016. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE; 17. EUHTACG. Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons (2024). https://health.ec.europa.eu/latest-updates/methodological-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons-2024-03-25_en; 18. Campbell & Gustafson 2022 American Statistician, 77(3): 248-258.

Presented at ISPOR EU Barcelona November 2024