Post-hoc validation of robust mixture priors for Bayesian control arm augmentation

A statistical test using marginal likelihoods

Daniel J. Sharpe[1], Tuli De[2], Jackie Vanderpuye-Orgle[2]

[1]Parexel International, London, UK
[3]Parexel International, Billerica, Massachusetts, USA

MSR166

Background

- Bayesian dynamic borrowing with robust mixture priors[1-3] is a recommended method for control arm augmentation[4] since it allows adaptive[5] leveraging of historical control data

Objectives

- To propose a statistical test for post-hoc validation of a prespecified weight for the informative component of a mixture prior distribution in an augmented control arm

Methods

Proposed test for mixture weight validation

- The prespecified weight w in a mixture prior can be validated post hoc via the ratios of marginal likelihoods[6] (i.e., the Bayes factors) when comparing to the model with the weight at the tipping point, at which the treatment effect is no longer statistically significant, and to the naïve model (i.e., $w = 0$)

Synthetic trial data and Bayesian borrowing model

- Synthetic data were generated for a randomized phase III trial of second- vs first-generation tyrosine kinase inhibitors in ROS1-positive advanced non-small cell lung cancer[7] (N=80, 3:1 randomization ratio), and for a historical trial of the control therapy (N=70). Both the current and historical studies had 36 months minimum follow-up duration
  - The primary endpoint in the phase III study was median progression-free survival (PFS), and the treatment effect was estimated using log-logistic distributions
- The informative components of the mixture priors for the parameters θ_c of the control arm model had a prespecified weight of $w = 0.6$ and were derived from a model fitted to historical (h) control data (estimates $\hat{\theta}_h$, variances σ_h),

$$\log \theta_c \sim w\text{Normal}\left(\hat{\theta}_h, \sigma_h\right) + (1-w)\text{Normal}(0,5). \quad (1)$$

Statistical goodness-of-fit: marginal likelihoods

- The marginal likelihood $\mathcal{Z}$ for a Bayesian model M with parameters $\boldsymbol{\theta}$ and prior distribution hyperparameters $\boldsymbol{\alpha}$, fitted to data $\mathcal{D}$,

$$\mathcal{Z}_M = \int p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\theta}, \quad (2)$$

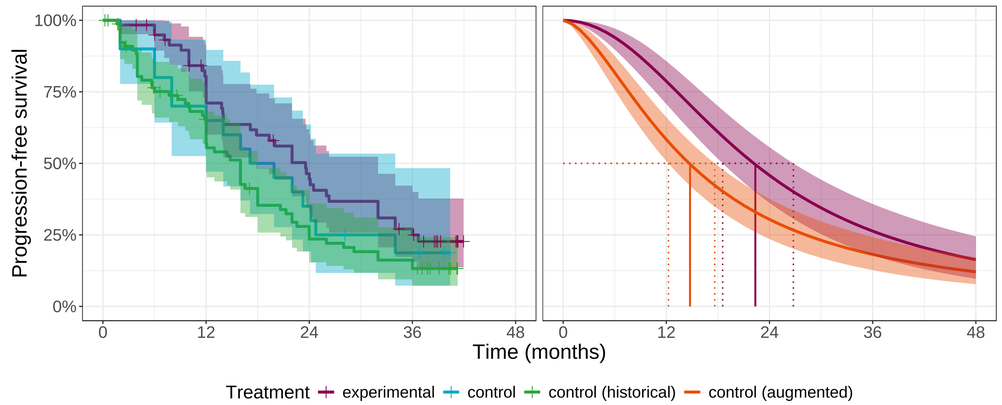
  is a rigorous measure of statistical goodness-of-fit[6,8]
- The odds ratio (Bayes factor[6]), $\mathcal{Z}_{M'}/\mathcal{Z}_M$ for model M' vs M, favors models with priors that have better overlap with the likelihood and do not have unnecessarily high uncertainty

Nested sampling

- The nested sampling algorithm,[9-11] which maintains a set of live points over a shrinking prior volume per an evolving likelihood threshold, presents an efficient and robust approach to calculate $\log \mathcal{Z}$ (cf. Eq. 2) and the approximate standard error thereof
- Marginal likelihoods and posterior samples were obtained by nested sampling with 60 live points, uniform sampling, and a tolerance of 0.01 for the approximate remaining contribution to $\log \mathcal{Z}$,[10,11] for models with w in increments of 0.1

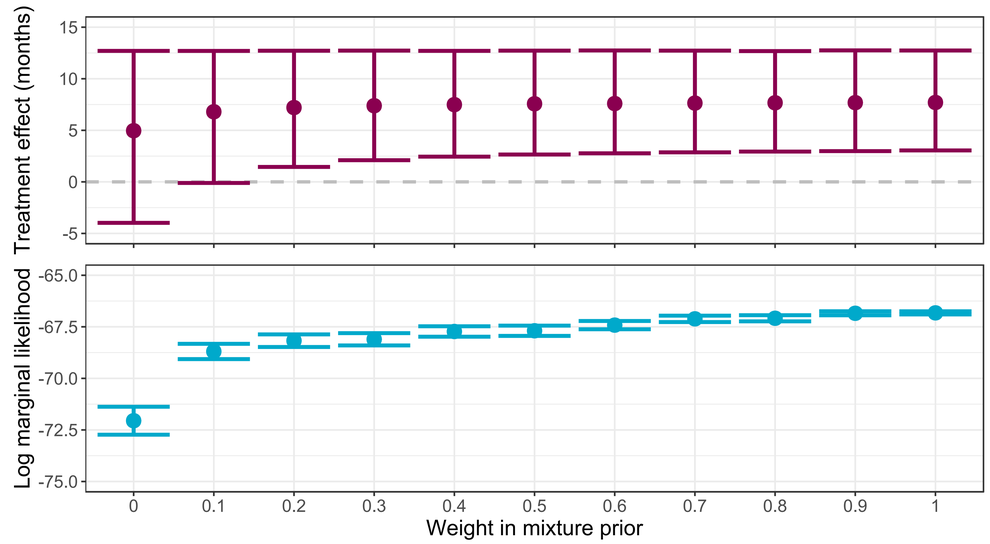
Results and Discussion

- Kaplan-Meier estimates for median PFS were 23.6 [95% CI: 17.9-32.0] vs 18.5 [12.0-34.0] months for experimental and control therapies, respectively, and 16.0 [12.0-18.0] months in the historical control data (Fig. 1)
- The treatment effect estimated from the prespecified Bayesian model ($w = 0.6$) was 7.6 [95% credible interval: 2.9 - 12.7] months, compared to 6.8 [-0.1 - 12.7] months in the tipping point model ($w = 0.1$) and 5.0 [-4.0 - 12.7] months in the naïve model ($w = 0$) (Fig. 2)
- Marginal likelihood increased with increasing weight of the informative component of the mixture prior (Fig. 2)
- The Bayes factors when comparing models with weights in the range $w = 0.6$ to 1 indicated no strong preference for any one model (odds ratios <2)
- The prespecified model ($w = 0.6$) was moderately favored over the model corresponding to the tipping point ($w = 0.1$, Bayes factor 4.9 [2.4-5.5]) and strongly favored over the naïve model ($w = 0$, Bayes factor 103.0 [67.9-156.2])
- The estimated Bayes factors support good exchangeability of the current and historical control data, thereby justifying the relatively strong borrowing in the prespecified model
- The metric $\mathcal{Z}$ penalizes excess prior uncertainty more strongly than the deviance information criterion or related measures, and is focused on model fit to observed data[6]



Shaded areas represent 95% uncertainty intervals for survival estimates, i.e., confidence and credible intervals in the Kaplan-Meier and Bayesian estimates, respectively. (Right) median PFS and 95% credible intervals are shown by vertical lines.

Figure 1: (Left) Kaplan-Meier estimates for PFS. (Right) Bayesian estimates for PFS when the control arm is augmented with historical trial data, using a mixture prior distribution with a weight of 0.6 for the informative component.



(A) Whiskers represent 95% credible intervals of the posterior distribution. (B) Whiskers represent approximate 95% confidence intervals estimated from the nested sampling algorithm.

Figure 2: (A) Treatment effect (difference in median PFS) estimated from Bayesian log-logistic survival models with mixture prior distributions, and (B) log marginal likelihood (goodness-of-fit, cf. Eq. 2) of the Bayesian models, with varying weight of the informative (historical control) component of the mixture priors.

Conclusions

- Estimation of marginal likelihoods for Bayesian augmented control arm models presents an interpretable approach for post-hoc validation of a prespecified mixture prior distribution, and may aid decision-makers to judge the reliability of studies borrowing from historical control data
  - Bayes factors can be used to clearly distinguish the suitability of the prespecified model compared to models that borrow less (e.g., the tipping point and naïve models) or more extensively from the historical control data
- Routine implementation of this test can lend further credibility to results from innovative trial designs in rare diseases

REFERENCES

[1] Viele K, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014; 13(1):41-54.
[2] Schmidli H, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; 70(4):1023-32.
[3] Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal*. 2012; 7(3):639-74.
[4] Ruberg SJ, et al. Application of Bayesian approaches in drug development: starting a virtuous cycle. *Nat Rev Drug Discov*. 2023; 22(3):235-50.
[5] Best N, et al. Assessing efficacy in important subgroups in confirmatory trials: an example using Bayesian dynamic borrowing. *Pharm Stat* 2021; 20(3):551-62.
[6] Pooley CM, Marion G. Bayesian model evidence as a practical alternative to deviance information criterion. *R Soc Open Sci*. 2018; 5(3):171519.
[7] Bischoff H, et al. Evolving therapeutic landscape of ROS1-positive non-small cell lung cancer: an updated review. *Curr Oncol*. 2025; 32(11):626.
[8] Del Moral P, Doucet A, Jasra A. Sequential Monte Carlo samplers. *J. R. Statist. Soc. B* 2006; 68(3):411-36.
[9] Skilling J. Nested sampling for general Bayesian computation. *Bayesian Anal*. 2006; 1(4):833-59.
[10] Higson E, Handley W, Hobson M, Lasenby A. Sampling errors in nested sampling parameter estimation. *Bayesian Anal*. 2018; 13(3):873-96.
[11] Higson E, Handley W, Hobson M, Lasenby A. Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. *Stat Comput*. 2018; 29:891-913.

© 2026 Parexel International (MA) Corporation

Contact: Daniel.Sharpe@parexel.com