A Multilevel Bayesian Model for Precision Oncology

ASHER WASSERMAN, xCures
JEFF SHRAGER, xCures
MARK SHAPIRO, xCures
AL MUSELLA, Musella Foundation for Brain Tumor Research and Information

The challenge of personalized medicine is to predict the causal effect of a treatment on patient, given a number of clinically relevant patient features. This task requires a flexible model that can integrate heterogeneous data, be easily interpreted by domain experts, and provide a meaningful quantification of the uncertainty in the prediction. In this submission we describe such a tool in the form of a multilevel Bayesian model for precision oncology, implemented in the probabilistic programming language, Stan, and we discuss the application of the model to brain cancer patient treatment outcomes.

1 INTRODUCTION

The problem of precision oncology – that is, precision medicine in cases of advanced cancer – is this: Given a description of a specific patient's state, including their history, and all available knowledge and data, choose the treatment that is most likely to have the greatest utility with respect to this patient's treatment goals. We here take the patient's goals to be prolonging of expected time to disease progression or death, as well as avoiding serious adverse events.

Physicians engaged in precision oncology must integrate an overwhelming amount of information from publications, and from their own experience. As of the end of 2019, PubMed reports 19,748 publications matching the term "breast cancer" in the past year alone, and the same search for open, recruiting studies in ClinicalTrials.gov returns 1,937 studies. Oncologists fighting less common cancers are in potentially a worse situation; Instead of being overwhelmed, they have only a few relevant publications, and may have seen only a small number of similar cases.

In this submission we describe a model whose purpose is to help oncologist predict outcomes for particular patients under different treatment regimens. Our goal for the model is to condition it on individual patient outcomes and/or summary statistics from clinical trials. Once conditioned, the model can predict outcomes for new patients under different treatment choices and provide a measure of the uncertainty of these predictions. The model's structure bears an understandable relationship to the domain, and to the types of inputs and outputs oncologists would expect. This may help users of the model to understand how the predictions, and uncertainty, are derived. We implement the model in the probabilistic programming language, Stan [Carpenter et al. 2017; Gelman et al. 2015].

2 DATA

We consider two types of patient outcome data that serve as the output of the generative model:

(1) Tumor load (TL), a measure or proxy of the volume of a patient's tumor
(2) Progression-free survival (PFS), the length of time after starting treatment that a patient lives without their disease getting worse

Tumor load is a longitudinal outcome, usually measured in regular cadences, and so for each patient we have a time series of TL measurements. For progression-free survival, we have a survival time and a binary indication of whether the failure event (disease progression or death) was

Authors' addresses: Asher Wasserman, xCures, 1901 Harrison Street, Suite 1100, Oakland, CA, 94612, awasserman@xcures.com; Jeff Shrager, xCures, jshrager@xcures.com; Mark Shapiro, xCures, mshapiro@xcures.com; Al Musella, Musella Foundation for Brain Tumor Research and Information.

observed by that time or was right-censored. There are a number of motivations for modeling these types of patient outcomes.

Tumor load, alone is an imperfect biomarker for survival, but when tumor location is included, this becomes a strong predictor for survival and other outcomes that are obviously important to patients, such as pain and other types of discomfort and functionally disabling symptoms that all reduce quality of life. Similarly, tumor load when considered as rate of change in tumor volume, when controlled for location, often provides a good measure of treatment activity. In most cases, when the tumor growth rate is reduced or reversed in temporal association with treatment, that is evidence of a response to the treatment.

Time-to-progression is a common outcome in clinical trials, which provides some interpretability of response measures in relation to clinical trials. Importantly time-to-progression does not always correlate with overall survival, especially in clinical trials where crossovers and subsequent therapies make post-progression analysis and interpretability more difficult. Another advantage of this model is that patients are observed over multiple lines of therapy in which prior lines of therapy and response become additional patient features. This facilitates the exploration of the relative contribution of different lines and interaction between lines of therapy to survival.

To predict patient outcomes, we use make use of two kinds of feature data: treatments and biomarkers. Patient can take a number of possible treatments at measured times, and so the TL time series data can be converted to a time offset from a start time of the first treatment. For this context we consider a treatment to be a monotherapy, with combinations of therapies represented as interaction terms in the treatment indicator vector, $\boldsymbol{x}_{\mathrm{tx}}$. Each patient also has a set of biomarkers, $\boldsymbol{x}_{\mathrm{bm}}$, that can be used to predict the utility of various treatments. Here we are using a very general definition of the term biomarker to include any clinically relevant features (genetic mutations, age, sex, tumor location, previous treatments, etc.). In principle these biomarkers can be time-dependent and continuous, but here we consider the simplification of static biomarkers with binary values (present or not present). For treatments that are assumed to have a direct causal interaction with a biomarker (e.g., a therapy targeted for a particular genetic mutation), we include an additional treatment-biomarker interaction term, $\boldsymbol{x}_{\mathrm{int}}$ which is equal to one for patients with that biomarker who were given the associated treatment.

3 MODEL

Many studies [e.g., Adrion and Mansmann 2012; Brown et al. 2005; Henderson 2000; Hickey et al. 2018; Król et al. 2018; Meller et al. 2019; Rizopoulos et al. 2014] have proposed and evaluated flexible longitudinal outcomes models, including joint models with time-to-event data. Here we model each type of outcome as a multilevel generalized linear response [Gelman and Hill 2007]. Population-level slope and intercept parameters are denoted as $\boldsymbol{\beta}$, while patient-level effects are denoted as $\boldsymbol{u}_i$. In the terminology of mixed models, these are referred to as fixed and random effects, respectively. A graphical summary of the generative model is shown in Figure 1.

3.1 Tumor load sub-model

For the i-th patient, the slope of the tumor load linear response is given by

$$\delta_i^{\mathrm{TL}} = \boldsymbol{x}_i^{\mathrm{TL}} \cdot \boldsymbol{\beta}^{\mathrm{TL}} + \boldsymbol{z}_i^{\mathrm{TL}} \cdot \boldsymbol{u}_{1i}^{\mathrm{TL}} \tag{1}$$

where $\boldsymbol{x}_i^{\mathrm{TL}} = (1, \boldsymbol{x}_{\mathrm{bm}_i}, \boldsymbol{x}_{\mathrm{tx}_i}, \boldsymbol{x}_{\mathrm{int}_i})$ is the vector of predictors, $\boldsymbol{z}_i^{\mathrm{TL}}$ is a subset of those predictors used for modeling patient-level effects, $\boldsymbol{\beta}^{\mathrm{TL}}$ is the vector of population-level effect sizes associated to the predictors, and $\boldsymbol{u}_{1i}^{\mathrm{TL}}$ are the patient-level slopes. For the j-th time series data point from the i-th patient, the mean linear response for tumor load is $\eta_{ij}^{\mathrm{TL}} = (\beta_0^{\mathrm{TL}} + u_{0i}^{\mathrm{TL}}) + \delta_i^{\mathrm{TL}}\ t_{ij}^{\mathrm{TL}}$ where β_0^{TL} is the

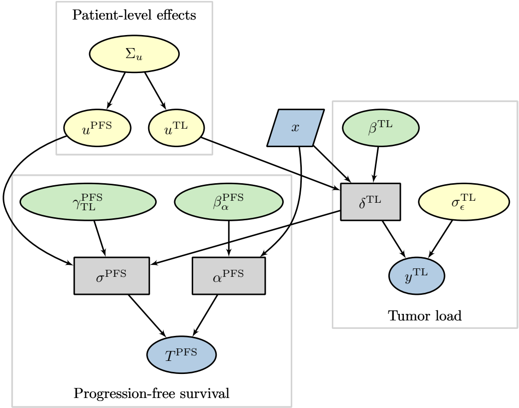


Fig. 1. Directed acyclic graph showing the causal assumptions in the model. Nodes with circles represent parameters whose values follow distributions conditioned on the values of parent nodes. Nodes with boxes represent parameters whose values are deterministically set by the value of parent nodes. Nodes colored blue indicate patient feature data (parallelogram) and patient outcome data (circles). Green nodes indicate population-level effect size parameters. Yellow nodes indicate sources of noise, both from measurement and from patient-level effects. For visual clarity, we omit intercept parameters.

population-level intercept, u^{TL}_{0i} is the patient-level intercept, and t^{TL}_{ij} is the time after treatment. We use a log-normal likelihood to restrict the model to positive TL values:

$$\log y^{\mathrm{TL}}_{ij} \sim \mathcal{N}(\eta^{\mathrm{TL}}_{ij}, \sigma^{\mathrm{TL}}_{\epsilon}) \tag{2}$$

3.2 Progression-free survival sub-model

We model the time-to-disease progression data using a log-logistic survival model, parameterized as an accelerated failure time model. The probability of having observed a disease progression time T_i for patient i for the log-logistic model is given by

$$\mathcal{L}^{\mathrm{PFS}}_{\mathrm{obs}}(T_i) = \frac{\alpha_i}{T_i}\left(\frac{T_i}{\sigma^{\mathrm{PFS}}_i}\right)^{\alpha_i}\left(1+\left(\frac{T_i}{\sigma^{\mathrm{PFS}}_i}\right)^{\alpha_i}\right)^{-2} \tag{3}$$

where $\sigma^{\mathrm{PFS}}_i = \exp\left(\gamma^{\mathrm{PFS}}_{\mathrm{TL}}\delta^{\mathrm{TL}}_i + \boldsymbol{z}^{\mathrm{PFS}}_i \cdot \boldsymbol{u}^{\mathrm{PFS}}_i\right)$ and $\alpha_i = \exp\left(\boldsymbol{x}^{\alpha}_i \cdot \boldsymbol{\beta}_{\alpha}\right)$ are the scale and shape parameter, respectively, of the log-logistic model. The parameter $\gamma^{\mathrm{PFS}}_{\mathrm{TL}}$ determines how a change in TL maps to a probability of the patient's disease being classified as progressive. $\boldsymbol{z}^{\mathrm{PFS}}_i$ is a vector of patient-level survival predictors while $\boldsymbol{u}^{\mathrm{PFS}}_i$ are the patient-level effects on survival. $\boldsymbol{x}^{\alpha}_i$ is a vector of predictors for survival curve shape, while $\boldsymbol{\beta}^{\alpha}$ are the population-level effects of the predictors on the survival curve shape. In the context the survival analysis literature, the exponential of the patient-level slope, $\exp u$, is occasionally referred to as frailty and allows for one to account for unobserved sources of variation in survival times [Keiding et al. 1997; Lambert et al. 2004].

Fortunately, many patients have right-censored survival times. The likelihood for observed survival times, $\tilde{T}_i$, when considering right-censored data is given by the log-logistic survival function,

$$\mathcal{L}^{\mathrm{PFS}}_{\mathrm{censored}}(T_i) = S(T_i) = \left(1+\left(\frac{T_i}{\sigma^{\mathrm{PFS}}_i}\right)^{\alpha_i}\right)^{-1} \tag{4}$$

3.3 Patient-level effects

The patient-level effects, $\boldsymbol{u}_i$, represent unmeasured sources of variation in outcomes. In this sense, they account for unknown confounders in the true data generating model. As described in the above subsections, each patient has multiple patient-level effect parameters. To handle the additional freedom this introduces into the model, we use an informative prior distributions over these effects. In particular, we assume that the patient-level effects, have a multivariate normal distribution such that

$$\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_u) \tag{5}$$

where Σ_u is the covariance matrix that represents latent associations between all patient-level random effects, and

$$\boldsymbol{u}_i = (u_{0i}^{\mathrm{TL}}, \boldsymbol{u}_{1i}^{\mathrm{TL}}, \boldsymbol{u}_i^{\mathrm{PFS}}) \ . \tag{6}$$

In practice we decompose Σ_u as

$$\Sigma_u = D_u \Omega_u = D_u L_u L_u^T \tag{7}$$

where

$$D_u = \mathrm{Diag}(\sigma^2_{u_0^{\mathrm{TL}}}, \sigma^2_{u_1^{\mathrm{TL}}}, \sigma^2_{u^{\mathrm{PFS}}}) \tag{8}$$

is the diagonal matrix of patient-level variance terms, Ω_u is the patient-level correlation matrix, and L_u is the lower triangular Cholesky decomposition of the correlation matrix, Ω_u.

4 PRACTICAL CONSIDERATIONS

An example Stan implementation of the survival sub-model is provided in the supplemental material. We use a wide normal prior for effect size parameters (e.g., $\boldsymbol{\beta}^{\mathrm{TL}}$, $\boldsymbol{\beta}^{\alpha}$) centered at 0. For noise scale parameters (e.g., $\boldsymbol{\sigma}_u$, $\sigma_\epsilon^{\mathrm{TL}}$), we follow [Gelman 2006] in using a weakly informative Half-Cauchy prior distribution.

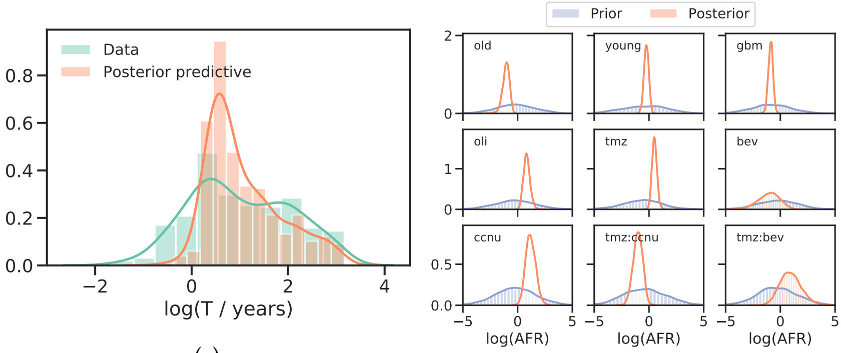
To code the patient feature vectors and build design matrices, we use *patsy*[1], a Python package for describing statistical models. Within the Stan code, we allow for any subset of population-level predictors to be used as patient-level predictors by passing in a vector of binary indicators of the same length as the feature vector. For choosing patient-level random effects, we start by assuming $z_i = 1$ (i.e., that the patient-level noise is uncorrelated with patient-features), then we introduce more patient-level predictors as warranted by significant improvements in the leave-one-out cross-validation (LOO-CV) information criterion [Vehtari et al. 2017].

We sample from the posterior probability distribution of the model using the No-U-Turn Hamiltonian Monte Carlo implementation in Stan [Hoffman and Gelman 2014], and we use the *CmdStanPy*[2] interface to call Stan from Python. For representing and storing the outputs of inferences with the model, we use *arviz*[3] [Kumar et al. 2019], a Python package for exploratory analysis of Bayesian models.

5 RESULTS

We condition the model as described in Section 3 on individual treatment outcomes for 362 patients with high-grade gliomas from the Musella Foundation Virtual Trial Registry[4]. This is a self-selected and self-reported observational dataset, and thus it represents a challenge for inferring causal effects of treatments. To predict treatment response, we use biomarkers of patient age at diagnosis, tumor type (one of glioblastoma multiforme (GBM), anaplastic astrocytoma, or oligodendroglioma),

[1] https://patsy.readthedocs.io/en/latest/index.html
[2] https://cmdstanpy.readthedocs.io/en/latest/index.html
[3] https://arviz-devs.github.io/arviz/index.html
[4] https://virtualtrials.com/brain/



(a)

(b)

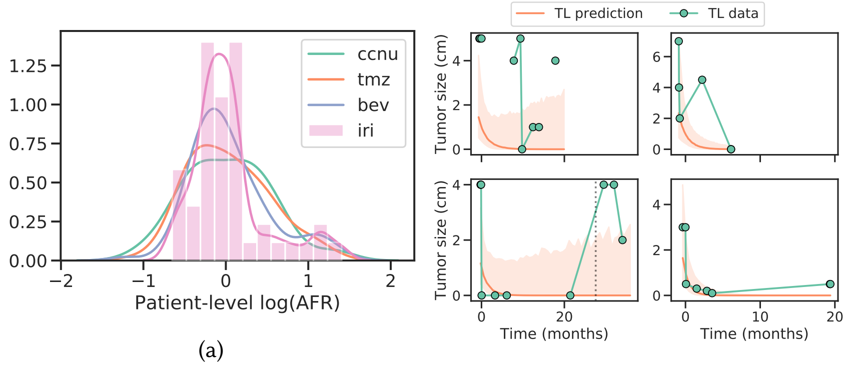
Fig. 2. Left panel: Posterior predictive check for survival times. Note that the posterior predictive survival times have been right-censored such that for patients with censored survival times, $\hat{T}_{i,\mathrm{obs}} = \min(\hat{T}_i, T_{i,censored})$. While the model recovers the skewed distribution of $\log(T)$, it under-predicts the shortest survival times. Right panels: Posterior (orange histograms) and prior (blue histograms) distributions for the effect size (as log accelerated failure rate) of different biomarkers and treatments. The effects (top to bottom, left to right) are for indicators of age $>$ 58 years, age $<$ 35 years, disease classified as glioblastoma, disease was classified as oligodendroglima, treated with temozolomide, treated with bevicizumab, treated with lomustine, treated with a combination of temozolomide and lomustine, and treated with a combination of temozolomide and bevacizumab. We see strong negative effects on survival of old age and GBM, while we see strong positive effects on survival from oligodendroglioma, temozolomide and lomustine.. There is some evidence for a positive effect from the interaction of bevacizumab and temozolomide.

and whether or not the tumor was resectable (i.e., able to be removed via surgery). We investigate 14 different treatments contained in the database. Many of these treatments are cytotoxic chemotherapies, though some of these treatments are intended as adjuvant therapies that are used to supplement other treatments. For the cases of combination therapies, we introduce pairwise interaction terms in the treatment predictors, $\boldsymbol{x}_{\mathrm{tx}}$.

As a test of the feasibility of extrapolating with this model, we initially restrict our inference to the survival data, then use the inferred posterior distributions to predict the tumor load trajectories. To do this, we set $\gamma_{\mathrm{TL}}^{\mathrm{PFS}} = 1$, $z_{1,i}^{\mathrm{TL}} = 0$, and $z_i^{\mathrm{PFS}} = 1$. The patient-level random effects in the model thus represent the combined (unobserved) confounders on both tumor load and progression-free survival.

Figure 2a compares the distribution of median posterior predictive survival times with that of the observed survival times. To facilitate the comparison, we right-censor any posterior predicted times associated to patients with censored survival times, such that $\hat{T}_{i,\mathrm{obs}} = \min(\hat{T}_i, T_{i,\mathrm{censored}})$. Figure 2b shows how the population-level (i.e., explained by the measured covariates) accelerated failure rate parameter distributions change after conditioning on the data.

By examining the distribution of patient-level effects split on patient covariate groups, we can look for clues that point toward unmeasured sources of variation in patient outcomes. For each treatment, we compute the posterior median patient-level effect size distribution for the subset of patients who were exposed to that treatment. We compute the p-value from the Shapiro-Wilk normality test, and find the lowest p-value (4×10^{-5}) for irinotecan, whose distribution of patient-level effects (along with those of other treatments) is shown in Figure 3a. The cluster of large patient-level effects in this treatment (as well as in bevacizumab) suggests that there may be



(a)

(b)

Fig. 3. Left panel: Patient-level effects (the set of $\{u_i\}$ from Equation 6, as log accelerated failure rates) across different subsets of patients based on treatment received. The treatments (top to bottom in legend, in order of most normal to least normal following the Shapiro-Wilks test p-value) are lomustine, temozolomide, bevacizumab, and irinotecan. We see that patient-level effects distributions of the irinotecan- and bevacizumab-selected patients have notable clusters of with larger effect sizes (associated to longer survival times). Right panel: Tumor size over time for four patients from the Musella Virtual Trial database, along with the posterior predictive distribution for TL as conditioned on the survival data. The green circles show the measured tumor sizes while the orange line shows the median posterior predictive TL. The shaded orange region show the 16th-84th percentile Bayesian credible interval. The vertical dashed line in the lower left panel shows the time of disease progression. We see that the model extrapolation capture the overall trends in the TL longitudinal data.

subgroups of patients who would benefit more from these treatments than the population as a whole. Of course, verifying that this is indeed a causal effect of the treatment is difficult to do outside of a randomized controlled trial.

Finally, we can use the model, conditioned on only the survival data, to predict what the longitudinal trajectories of tumor size should look like. Figure 3b shows these predictions for four patients with longitudinal measurements of tumor sizes. To make these predictions, we assume $\sigma^2_{u^{\mathrm{TL}}} = 0$, and $\sigma_\epsilon^{\mathrm{TL}} = 1$, but we emphasize that these predicted TL trajectories are extrapolations of the model as conditioned on the survival data, not fits to the TL data themselves.

6 CONCLUSIONS AND FUTURE WORK

In this submission, we have demonstrated the application of a Bayesian multilevel model for cancer patient treatment outcomes. Further development of the model will focus on modeling multiple types of longitudinal patient outcomes (e.g., patient functional performance scores and other proxies for tumor load) and rates of serious adverse treatment effects. We are exploring the application of the model to clinical trial results, which will help ensure the robustness of any causal inferences made with the model.

REFERENCES

Christine Adrion and Ulrich Mansmann. 2012. Bayesian Model Selection Techniques as Decision Support for Shaping a Statistical Analysis Plan of a Clinical Trial: An Example from a Vertigo Phase III Study with Longitudinal Count Data as Primary Endpoint. *BMC Medical Research Methodology* 12, 1 (December 2012), 137. https://doi.org/10.1186/1471-2288-12-137

Elizabeth R. Brown, Joseph G. Ibrahim, and Victor DeGruttola. 2005. A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival. *Biometrics* 61, 1 (2005), 64–73. https://doi.org/10.1111/j.0006-341X.2005.030929.x

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76, 1 (January 2017), 1–32. https://doi.org/10.18637/jss.v076.i01

Andrew Gelman. 2006. Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper). *Bayesian Analysis* 1, 3 (September 2006), 515–534. https://doi.org/10.1214/06-BA117A

Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge ; New York. OCLC: ocm67375137.

Andrew Gelman, Daniel Lee, and Jiqiang Guo. 2015. Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics* 40, 5 (October 2015), 530–543. https://doi.org/10.3102/1076998615606113

R. Henderson. 2000. Joint Modelling of Longitudinal Measurements and Event Time Data. *Biostatistics* 1, 4 (December 2000), 465–480. https://doi.org/10.1093/biostatistics/1.4.465

Graeme L. Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. 2018. joineRML: A Joint Model and Software Package for Time-to-Event and Multivariate Longitudinal Outcomes. *BMC Medical Research Methodology* 18, 1 (December 2018), 50. https://doi.org/10.1186/s12874-018-0502-1

Matthew D Hoffman and Andrew Gelman. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15 (April 2014), 31.

Niels Keiding, Per Kragh Andersen, and John P. Klein. 1997. The Role of Frailty Models and Accelerated Failure Time Models in Describing Heterogeneity Due to Omitted Covariates. *Statistics in Medicine* 16, 2 (1997), 215–224. https://doi.org/10.1002/(SICI)1097-0258(19970130)16:2<215::AID-SIM481>3.0.CO;2-J

Agnieszka Król, Christophe Tournigand, Stefan Michiels, and Virginie Rondeau. 2018. Multivariate Joint Frailty Model for the Analysis of Nonlinear Tumor Kinetics and Dynamic Predictions of Death. *Statistics in Medicine* 37, 13 (2018), 2148–2161. https://doi.org/10.1002/sim.7640

Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. 2019. ArviZ a Unified Library for Exploratory Analysis of Bayesian Models in Python. *Journal of Open Source Software* 4, 33 (January 2019), 1143. https://doi.org/10.21105/joss.01143

Philippe Lambert, Dave Collett, Alan Kimber, and Rachel Johnson. 2004. Parametric Accelerated Failure Time Models with Random Effects and an Application to Kidney Transplant Survival. *Statistics in Medicine* 23 (2004).

Matthias Meller, Jan Beyersmann, and Kaspar Rufibach. 2019. Joint Modeling of Progression-free and Overall Survival and Computation of Correlation Measures. *Statistics in Medicine* 38, 22 (September 2019), 4270–4289. https://doi.org/10.1002/sim.8295

Dimitris Rizopoulos, Laura A. Hatfield, Bradley P. Carlin, and Johanna J. M. Takkenberg. 2014. Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging. *J. Amer. Statist. Assoc.* 109, 508 (October 2014), 1385–1397. https://doi.org/10.1080/01621459.2014.931236

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* 27, 5 (September 2017), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4 arXiv:1507.04544