Predicting Future OS behavior using Bayesian Posterior Predictive Distributions

Pourab Roy[1], Erik Bloomquist[1], Shiqiang Jin[2], Shenghui Tang[1]

[1]US Food and Drug Administration, [2]Kansas State University

Disclaimer: This work is based on the authors' personal opinions and does not reflect the official position of the FDA

1. Introduction

Time-to-event endpoints such as overall survival (OS) and progression-free survival (PFS) are commonly used to determine the efficacy of drugs in oncology. Usually efficacy is determined by performing hypotheses tests at pre-specified interim and final analyses. The results based on interim analyses often exhibit large variability and uncertainty due to immature data. A common clinical question of relevance is to forecast how these results will look with continued follow-up.

We adopted the use of Bayesian predictive probability to implement a prediction model, assuming an underlying piece-wise exponential distribution of the survival times, to simulate future behavior based on the current data. The goal of the approach is to check whether this method can predict the robustness of the interim OS analysis results from different approved clinical trials and determine whether they are representative of the final results.

To evaluate the approach, we present a comparison of the predicted and actual performance of six different clinical trial datasets. We have also developed a Rshiny app based on the methodology.

2. Methodology

The posterior predictive distribution is the distribution of the unobserved (censored) observations conditional on the observed data. Our goal is to predict the outcome of the final analysis based on the results from the interim analysis. Since the final analyses occurs sometime after the interim analysis, this essentially boils down to predicting the future behavior of the censored observations.

To achieve this goal, we assume that the control group follows a piecewise exponential distribution, with a hazard rate given by

$$\lambda_0(t) = \lambda_j, when\ t \in [T_j, T_{j+1}),$$

where the cut-points T_1, T_2,…., T_J are such that $0 = T_0 < T_1 < T_2 < \ldots < T_J < \infty$. We also assume that the Cox proportional hazards (PH) assumption is valid for the treatment and control groups. Thus, we can write the PH model as

$$\lambda_i(t|x) = \lambda_0(t)exp(\beta x),$$

where x=1 if the patient belongs to the treatment group and 0 otherwise.

Next, we assume a prior distribution on β and λ, i.e. $\beta, \lambda \sim p(\alpha)$. Then, for a given set of N observations Y=($Y_1, Y_2, \ldots, Y_N$), the posterior distribution is defined by the density function $p(\beta, \lambda | Y, \alpha) \propto p_F(Y|\beta, \lambda)p(\alpha)$, where $p_F(Y|\beta, \lambda)$ is the density function corresponding to $\lambda_i(t|x)$.

Thus, the posterior predictive distribution of a new observation $\tilde{y}$ can be calculated as

$p(\tilde{y}|Y, \alpha) = \int_{\beta,\lambda} p_F(\tilde{y}|\beta, \lambda)p(\beta, \lambda|Y, \alpha)d\beta d\lambda$.

The censored observations can thus be predicted using the following method: Use Gibbs sampling to estimate β, based on the given data at the interim analysis. Once an estimate of β is obtained, predict the i-th censored observation using $E(Y_i|Y_i > y_i)$. To mimic the final analysis, we have used two different approaches, a time-based approach and an event-based approach. In the event-based approach, we assume that the final analysis will be conducted after x% more events have occurred and the predictions are censored at the appropriate time point. Similarly, in the time-based approach, it is assumed that the final analysis will take place after x many months and all predictions are censored at that time point. In either approach, the entire process is repeated multiple times to get a set of final datasets.

For most of our results, we have assumed that J =1 or 2, with cut points at 3 months and/or 6 months and have used 100 resamples.

3. Results

To validate our method, we look at its performance for different data sets, under different sets of conditions. We look at the performance of the method for five different clinical trials, spanning three different cancer types – breast cancer, renal cell carcinoma and head and neck cancer. We have considered the underlying distribution to be piece-wise exponential with break points at 3 and 6 months and 100 resamples.

3.1 Event-Based Approach

For the first set of results, we have only considered the final analysis data and created the interim analysis datasets artificially, assuming the interim analysis occurred when 70%, 80% or 90% of the events were observed. The dataset was then truncated at the relevant time point to create the interim analysis data. The estimates were then created using the event-based approach, assuming an exponential distribution with break points at 3 and 6 months and using 100 resamples. The results are summarized in Table 1.

Table 1: Summary of Results from the Event-Based Analyses

Population	Observed p-value at IA (HR)	Estimated HR					Observed p-value at FA (HR)
		Median	Range	95% CI	# HR<0.8	# HR<0.9	
IA at 70% of the total # of events							
Trial A	0.140 (0.804)	0.878	(0.721, 1.075)	(0.757, 0.993)	17	64	0.142 (0.834)
Trial B	0.072 (0.704)	0.796	(0.625, 0.953)	(0.665, 0.914)	52	94	0.004 (0.625)
Trial C	0.637 (0.909)	0.976	(0.767, 1.319)	(0.810, 1.199)	2	22	0.156 (0.784)
Trial D	0.049 (0.668)	0.752	(0.594, 0.974)	(0.645, 0.928)	77	96	0.035 (0.693)
Trial E	0.000 (0.406)	0.522	(0.447, 0.610)	(0.475, 0.591)	100	100	0.000 (0.410)
IA at 80% of the total # of events							
Trial A	0.177 (0.830)	0.864	(0.761, 1.028)	(0.790, 0.952)	6	71	0.142 (0.834)
Trial B	0.028 (0.669)	0.733	(0.596, 0.880)	(0.653, 0.848)	81	100	0.004 (0.625)
Trial C	0.386 (0.848)	0.894	(0.696, 1.132)	(0.766, 1.063)	9	51	0.156 (0.784)
Trial D	0.046 (0.681)	0.731	(0.606, 0.857)	(0.620, 0.853)	83	100	0.035 (0.693)
Trial E	0.000 (0.396)	0.461	(0.406, 0.533)	(0.419, 0.512)	100	100	0.000 (0.410)
IA at 90% of the total # of events							
Trial A	0.062 (0.784)	0.808	(0.726, 0.874)	(0.743, 0.862)	44	100	0.142 (0.834)
Trial B	0.011 (0.644)	0.682	(0.604, 0.759)	(0.616, 0.755)	100	100	0.004 (0.625)
Trial C	0.228 (0.805)	0.811	(0.696, 0.936)	(0.739, 0.906)	31	96	0.156 (0.784)
Trial D	0.023 (0.662)	0.69	(0.597, 0.784)	(0.633, 0.769)	100	100	0.035 (0.693)
Trial E	0.000 (0.403)	0.445	(0.404, 0.489)	(0.410, 0.473)	100	100	0.000 (0.410)

We also wanted to come up with metrics to predict the efficacy at the final analysis. For this, we looked at the proportion of times the hazard ratio (HR) was less than 0.8 and 0.9 out of the 100 resamples. This metric can be tailored to an individual trial, based on the details in the protocol. Based on the table, we can see that median HR is slightly higher than the observed HR at the final analysis in most cases, which means the results based our method are conservative. The range and 95% CI are wider when the interim analysis occurs

at 70% of the number of events for the final analysis, but it becomes narrower as this percentage increases and our predictions also improve in quality. This is expected, because we are predicting a smaller proportion of the data as the percentage of events at the interim analysis increases. In all cases, the observed HR at the final analysis lies within the 95% CI predicted by the method. A visual representation of the results for Trial A and Trial B are given in Figures 1 and 2 respectively. In each of the graphs, the blue lines represent the predicted treatment curves and the orange lines represent the predicted control curves. If the two sets of curves do not overlap too much, it indicates that there is some difference in the efficacy outcomes between the two arms, while a large amount of overlap signifies that there are similarities in the survival outcomes in the treatment and control group and a treatment effect may not be present. In both Figures 1 and 2 we see that the actual Kaplan-Meier curves at the final analysis (black lines) lie within the corresponding prediction bands, which shows that the method is performing as expected.

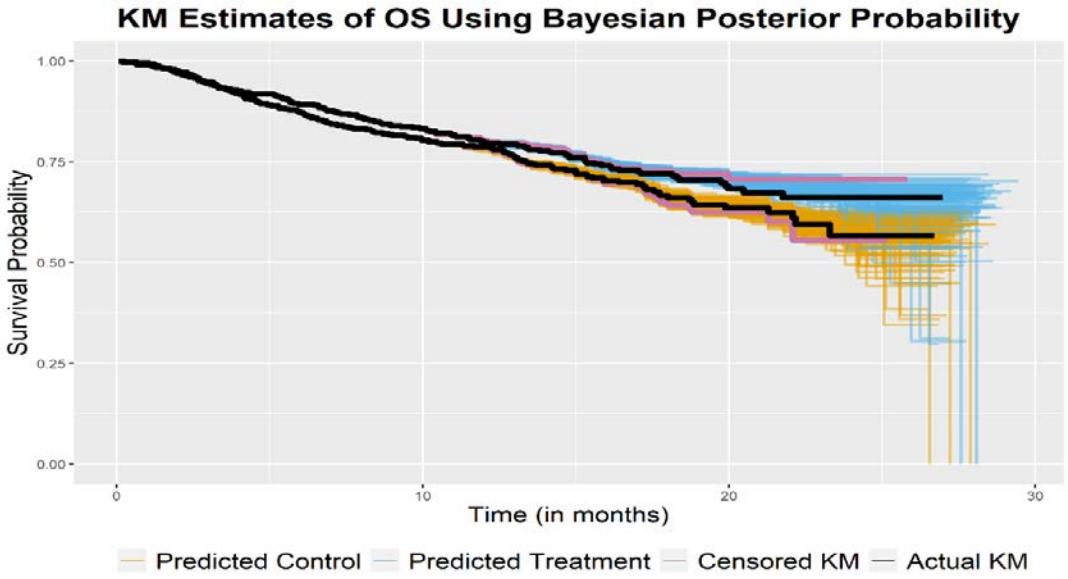

Figure 1: Kaplan-Meier Plot for Trial A when the IA occurs at 90% of the total events

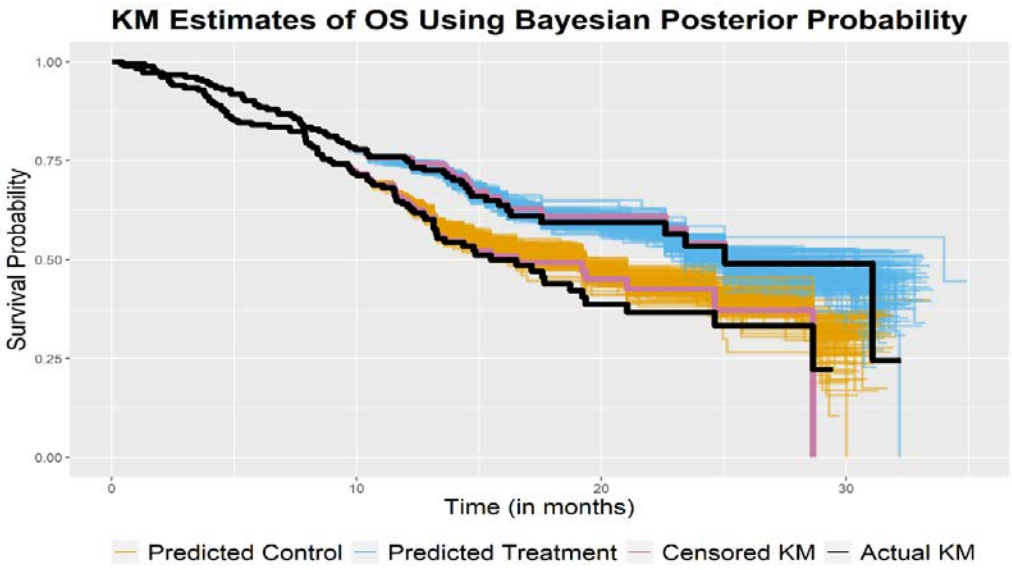

Figure 2: Kaplan-Meier Plot for Trial B when the IA occurs at 90% of the total events

3.2 Time-Based Approach

Next, we created the interim analysis datasets by assuming the interim analysis occurred 3 or 6 months before the final analysis. The dataset was then truncated at the relevant time point to create the interim analysis data. The estimates were then created using the time-based approach, assuming an exponential distribution with break points at 3 and 6 months and using 100 resamples. The results are summarized in Table 2.

Table 2: Summary of Results from the Time-Based Analyses

Population	Observed p-value at IA (HR)	Estimated HR					Observed p-value at FA (HR)
		Median	Range	95% CI	# HR<0.8	# HR<0.9	
IA 3 months before the FA							
Trial A	0.094 (0.802)	0.841	(0.752, 0.953)	(0.762, 0.903)	22	95	0.142 (0.834)
Trial B	0.018 (0.651)	0.690	(0.616, 0.798)	(0.621, 0.773)	100	100	0.004 (0.625)
Trial C	0.414 (0.856)	0.920	(0.761, 1.206)	(0.787, 1.074)	4	43	0.156 (0.784)
Trial D	0.021 (0.655)	0.691	(0.588, 0.787)	(0.613, 0.765)	100	100	0.035 (0.693)
Trial E	0.000 (0.403)	0.599	(0.471, 0.780)	(0.497, 0.729)	100	100	0.000 (0.410)
IA 6 months before the FA							

Trial A	0.166 (0.818)	0.868	(0.734, 1.078)	(0.756, 0.986)	13	73	0.142 (0.834)
Trial B	0.105 (0.717)	0.825	(0.670, 1.025)	(0.685, 0.980)	34	79	0.004 (0.625)
Trial C	0.090 (0.663)	0.871	(0.599, 1.194)	(0.632, 1.094)	30	63	0.156 (0.784)
Trial D	0.036 (0.651)	0.737	(0.607, 0.988)	(0.653, 0.878)	85	98	0.035 (0.693)
Trial E	0.016 (0.450)	0.904	(0.543, 1.339)	(0.646, 1.189)	25	48	0.000 (0.410)

Based on the results, we can see that the 3 months prediction are quite conservative and seems to be doing a good job at predicting the results. The 6-month predictions are actually quite unstable, with much wider 95% confidence intervals. This shows that at 6 months out, the predictions do not work very well and there is a lot of uncertainty around the interim analysis data. A visual representation for Trials A and B is given in Figures 3 and 4 respectively.

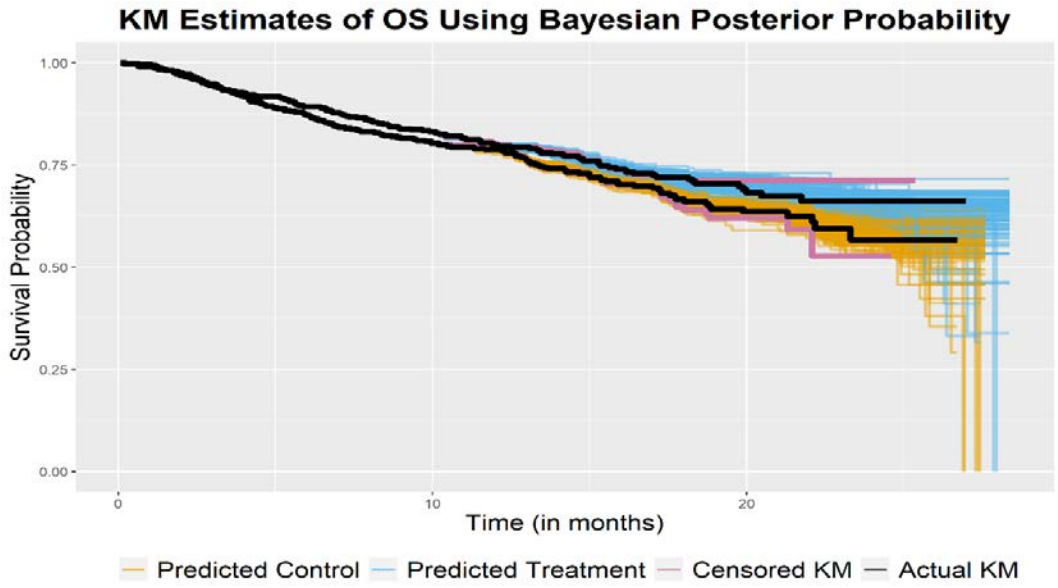

Figure 3: Kaplan-Meier Plot for Trial A when the IA occurs 3 months before the FA

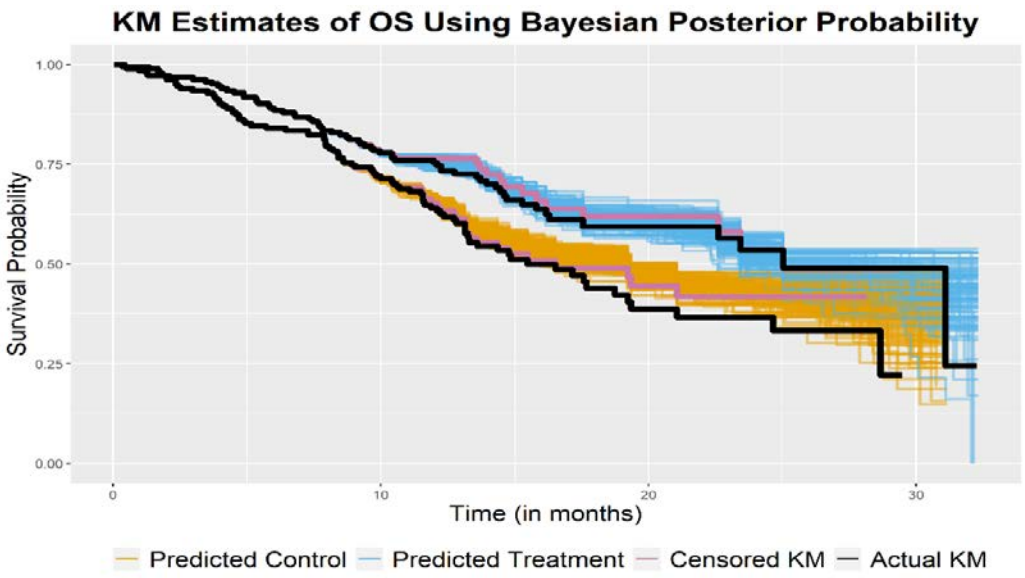

Figure 4: Kaplan-Meier Plot for Trial B when the IA occurs 3 months before the FA

3.1: Sensitivity

In this part, we look at the impact of the different model assumptions made during estimation. The first set of results looks at the impact of the piece-wise exponential assumption. We compare three different set-ups: piece-wise exponential with one break at 3 months, piece-wise exponential with one break at 6 months and piece-wise exponential with two breaks at 3 and 6 months. The dataset compared here is from Trial F, which has been approved for head and neck cancer. The trial looked at 6 different subpopulations and the results are compared for all 6 of them. We specifically looked at this dataset for this exercise because the exponential assumption does not hold for the underlying data and we felt that the impact of the piece-wise exponential assumption may be most prominent in this dataset.

Table 3: Impact of Piece-wise Exponential Assumption on the Results for Trial F

Population	Estimated Median P-values (Range)		
	One break at 3 months	One break at 6 months	Two breaks at 3 and 6 months
Subpopulation 1	0.032 (0.002, 0.131)	0.029 (0.002,0.144)	0.024 (0.001,0.216)
Subpopulation 2	0.003 (0.000,0.025)	0.003 (0.000,0.026)	0.003 (0.000,0.026)
Subpopulation 3	0.009 (0.001,0.049)	0.008 (0.000,0.062)	0.006 (0.000,0.051)
Subpopulation 4	0.016 (0.001,0.114)	0.012 (0.001,0.068)	0.012 (0.000,0.051)
Subpopulation 5	0.015 (0.001,0.094)	0.013 (0.001,0.083)	0.012 (0.000,0.080)
Subpopulation 6	0.054 (0.013,0.252)	0.058 (0.005,0.363)	0.067 (0.010,0.218)

Based on the above set of results, we can see that the impact of the number of breaks is minimal. Similar results were also observed for Trials A-E. Thus, we decided to use 2 break points at 3 and 6 months for our analyses.

The final set of results looks at the impact of the number of resamples. We compare the results obtained under 50, 100 and 200 resamples. A part of it has been summarized in Table 6.

Table 4: Impact of the Number of Resamples on the Results for Trial F

Population	Estimated Median P-values (Range)		
	50 resamples	100 resamples	200 resamples
Subpopulation 1	0.023 (0.001,0.113)	0.024 (0.001,0.216)	0.026 (0.001,0.216)
Subpopulation 2	0.004 (0.000,0.019)	0.003 (0.000,0.026)	0.003 (0.000,0.035)
Subpopulation 3	0.006 (0.000,0.051)	0.006 (0.000,0.051)	0.008 (0.000,0.051)
Subpopulation 4	0.014 (0.000,0.059)	0.012 (0.000,0.080)	0.012 (0.000,0.080)
Subpopulation 5	0.010 (0.001,0.083)	0.012 (0.001,0.083)	0.010 (0.001,0.084)
Subpopulation 6	0.071 (0.024,0.218)	0.067 (0.010,0.218)	0.070 (0.006,0.289)

Based on the above set of results, we can see that an increase in the number of resamples to 200 did not lead to any significant improvement in the results. Thus, we decided to use 100 resamples for all our calculations.

4. Discussion

Based on the results, the Bayesian posterior prediction model does well when predicting the behavior at the final analysis, based on the results at the interim analysis. The method seems to be conservative and works well when using the event-based approach. The method, however, breaks down when we predict more than 6 months out using the time-based analysis. In these cases, one must caution that although the model provides an estimate, they may reflect actual performance from previous studies. As an alternative, an event-based approach may allow for prediction at a pre-specified number of events, however predictions that occur into the distant future may still be unreliable

Our future goal is to apply the method to a larger number of trials to determine if there are specific situations when this method breaks down. We also hope to implement a more robust piece-wise continuous function that does not rely on a proportional hazards assumption. We would also like to look at the model performance under such conditions.