



Model-Based Analysis of Disease Progression in Pancreatic Cancer Based on Registry Data Collected in the Surveillance Epidemiological End Result (SEER) Database

Gaurav Bajaj¹, Erin Dombrowsky¹, Qilu Yu², Shashank Rohatagi³, Jeffrey S. Barrett¹

¹Laboratory of Applied PK/PD, Department of Clinical Pharmacology and Therapeutics, The Children's Hospital of Philadelphia, PA, USA; ²Westat Inc., Rockville, MD, USA; ³ Piramal Life Sciences, Mumbai, India

Objective

- Develop a parametric survival model using data from Surveillance, Epidemiology, and End Result (SEER) registry database¹, identify relevant covariates to stratify patients in future trials, and predict disease outcome at initial stages.
- Evaluate the suitability of model based on SEER data relative to historical experience and pancreatic cancer trials from two different institutions.

Clinical presentation of Pancreatic Cancer

- With most patients diagnosed with pancreatic adenocarcinoma (PAC) in advanced stages and few months to survive, the disease is considered largely incurable due to minimal effects of treatment and overall survival in patients after diagnosis is 5-6 months^{2,3}.
- Standard treatment includes chemo-/radio-therapy in case of non-resectable tumors, surgery followed by chemo-/radio-therapy for resectable tumors (10-20% cases).
- Recent trials with combination chemotherapy show a fractional increase in the life time survival of patients^{2,3}.
- Sample size in pancreatic cancer (PC) trials is usually small and poses significant problem to compare treatment across trials.

Dataset for modeling Pancreatic Cancer

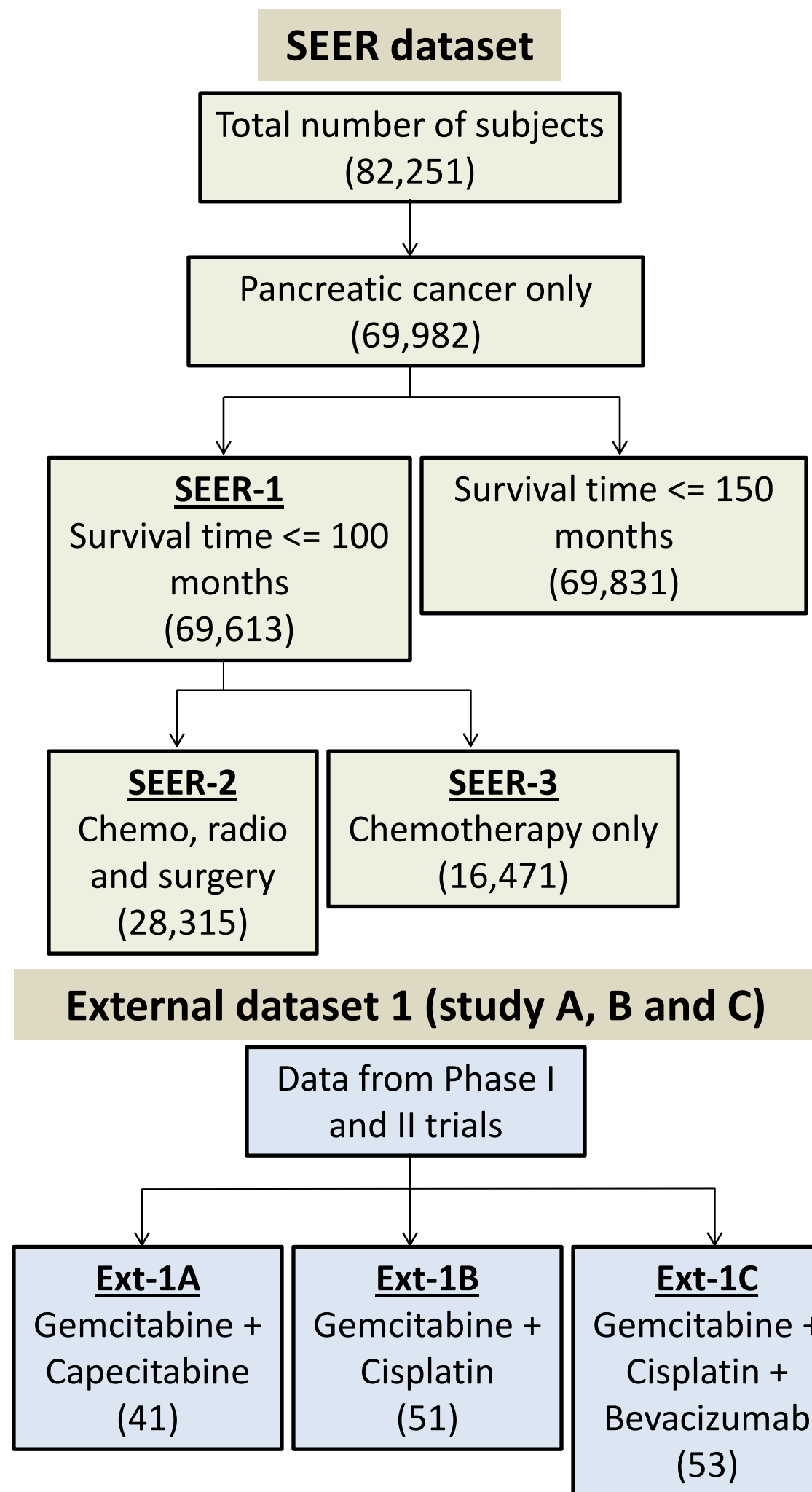


Figure 1. Datasets used in modeling Pancreatic Cancer . PC data was extracted from SEER website and refined based on disease status and survival time. Data was further classified based on treatment. External dataset (received from two different institutions) were used for model validation. Number of subjects in each dataset are shown in parenthesis. None of the dataset had longitudinal information on disease or treatment. Datasets used for analysis include: SEER-1, 2, 3; Ext-1A, 1B, 1C; Ext-2A, 2B.

Non-parametric analyses

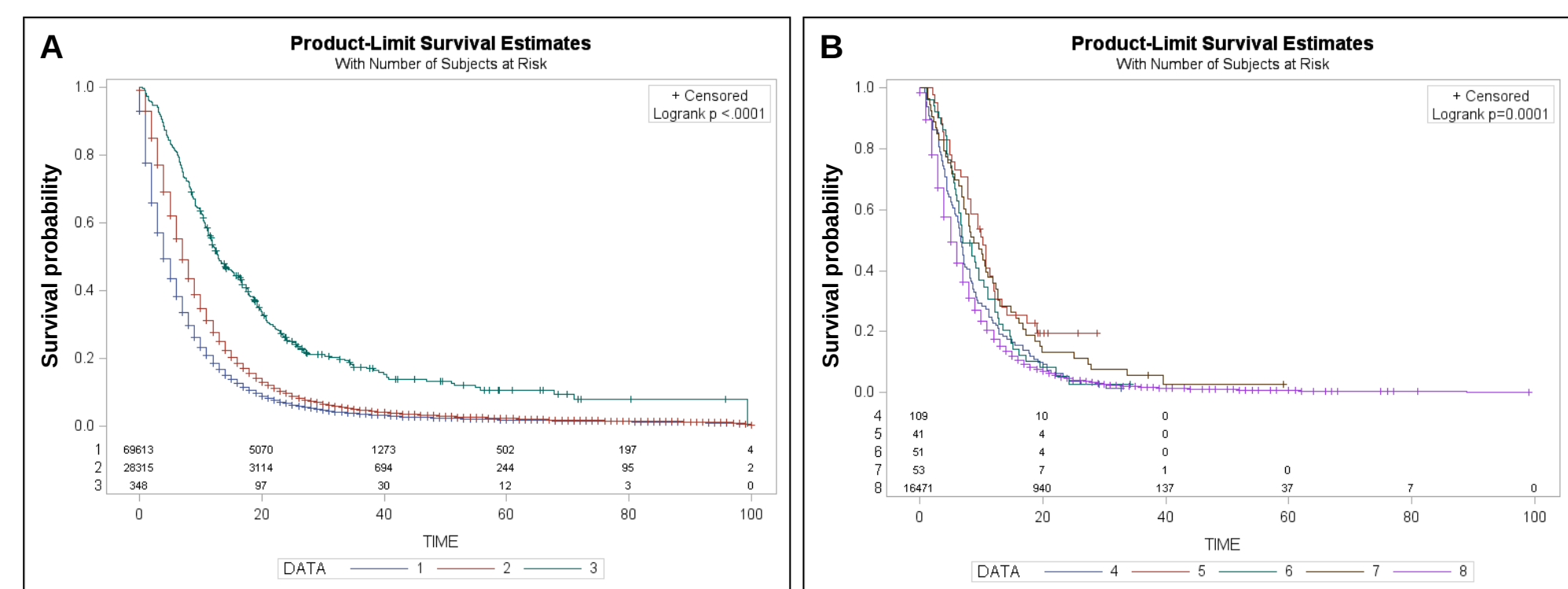


Figure 2. Non-parametric analyses of SEER and External datasets used in modeling. A) Kaplan-Meier (KM) curves for SEER and external dataset with chemo- (with/without radiotherapy and surgery). B) KM curves of dataset with chemotherapy only. X-axis shows the survival time in months. Data # and median survival time are shown in the table below.

Table 1. Median survival time of SEER and External datasets

	SEER-1	SEER-2	Ext -2A	Ext -2B	Ext-1A	Ext-1B	Ext-1C	SEER-3
DATA #	1	2	3	4	5	6	7	8
Median survival time (months, SE)	4	7	12.94 (11.79 – 15.47)	6.73 (5.74 – 7.62)	10.2 (7.9- (12)	7.07 (5.82- 9.67)	8.78 (6.9- 11.84)	5
Chemo (C), radio (R) and surgery (S)	C R S (or no therapy)	C (R S)	C (R S)	C (R S)	Chemotherapy only (C)			

Semi parametric analyses using SEER data

Table 2. Information on Covariates in SEER data

Type	Predictor variables
Tumor Characteristics	Tumor size, Lymph node status, Tumor Stage, Grade, Histology, Site
Therapy	Surgery , Radiation, Chemotherapy
Patient demographics	Race, Gender, Age
Age, Size, Regional lymph node status are continuous predictors, all other are categorical predictor variables.	

Limitations of SEER dataset:

- SEER dataset includes chemotherapy (chemo) as categorical variable: 1- patients who received chemo, 0-without chemo. Patients under without chemo subcategory also include patients with unknown information on chemo. In addition, SEER data does not include information on drugs used for chemo.
- No longitudinal information on disease or treatment.
- One of the important predictors of PAC is obstructive jaundice and abdominal pain which is not in SEER dataset.

$$f(t) = S(t)h(t)$$
$$H(t) = \int_0^t h(x)dx$$
$$S(t) = \exp\left(-\int_0^t h(x)dx\right)$$

Cox proportional hazard model:

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$$

$h_0(t, X) = e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$ HR = hazard ratio

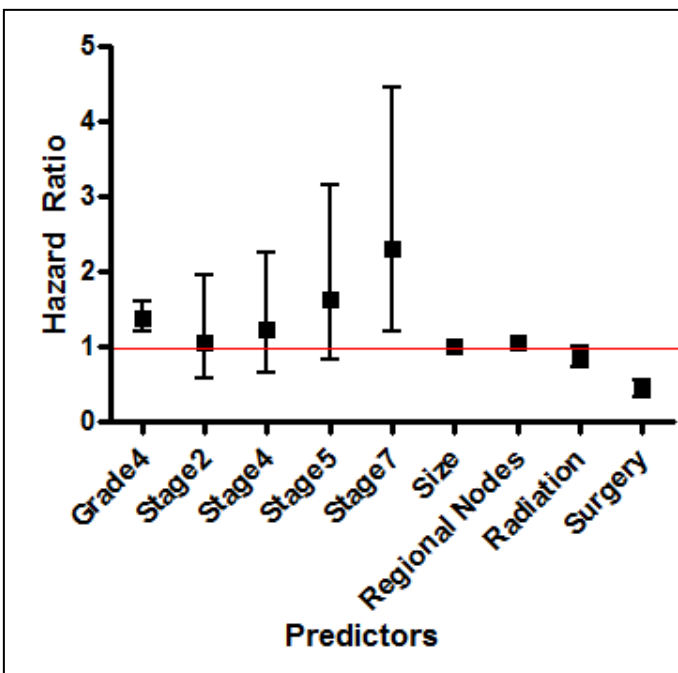


Figure 3. Hazard ratio (HR) for semi-parametric model. Figure shows hazard ratio of predictors from a semi-parametric model. HR <1 indicates decrease in hazard and HR>1 increased hazard. For tumor grade and stage, the levels were compared to grade/stage level =1.

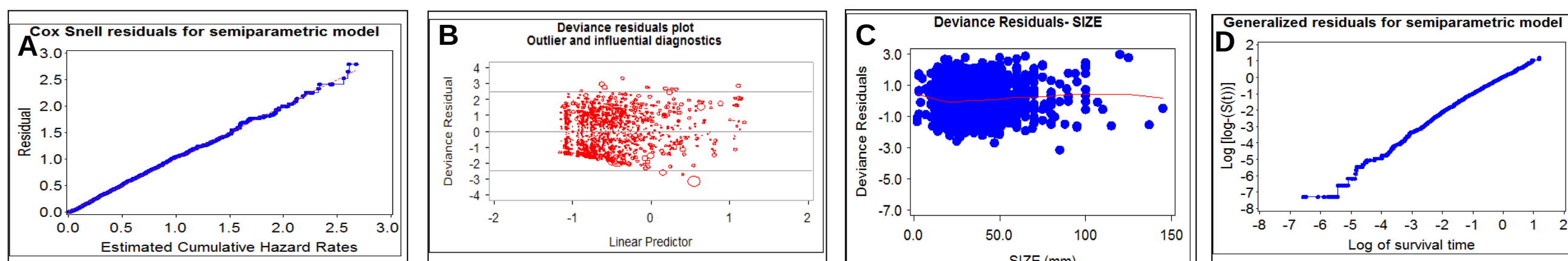


Figure 4. Diagnostic plots for evaluation of semi-parametric model. A) Cox-snell residual for overall fit of model, B and C) Model evaluation using deviance residuals with linear predictor and continuous covariate (tumor size) for model evaluation, D) Generalized residual plot to check model fit. Figures show that semi-parametric model was fitted well. Proportional hazard assumptions were evaluated using martingale residuals (data not shown).

Parametric analyses using SEER data

$$Y = \log(T_i) = \beta_0 + \beta_1 X_i + \sigma \epsilon$$

OR

$$T_i = \exp(\beta_0 + \beta_1 X_i) + \exp(\sigma \epsilon)$$

$\log(T_i)$ = log of a survival time, β = vector of model parameters corresponding to the covariate vector X_i , ϵ = random error that follows an extreme value distribution, σ = scale parameter

For gamma distribution: $f(t) = \frac{\lambda p(\lambda t)^{p-1} e^{-(\lambda t)^p}}{\Gamma(p)}$

For log-logistic distribution: $S(t) = \frac{1}{1 + (\lambda t)^p}$ $\lambda(t) = \frac{\lambda p(\lambda t)^{p-1}}{[1 + (\lambda t)^p]^2}$

where, $f(t)$ = probability density function, $\lambda(t)$ is hazard function, $p = 1/\epsilon$

The above statistical model was used to fit the log of survival time as a linear function of covariates. Pancreatic Cancer data was fitted to different parametric distributions that are commonly used in survival analysis. Model was then selected based on the fitting criteria and residual analysis.

Type of model	AIC (SEER-2)	AIC (SEER-1)
Exponential	2810.19	5358.034
Weibull	2602.66	5238.03
Log-logistic	2474.29	5144.97
Log normal	2448.72	5320.819
Gama	2429.29	5189.97

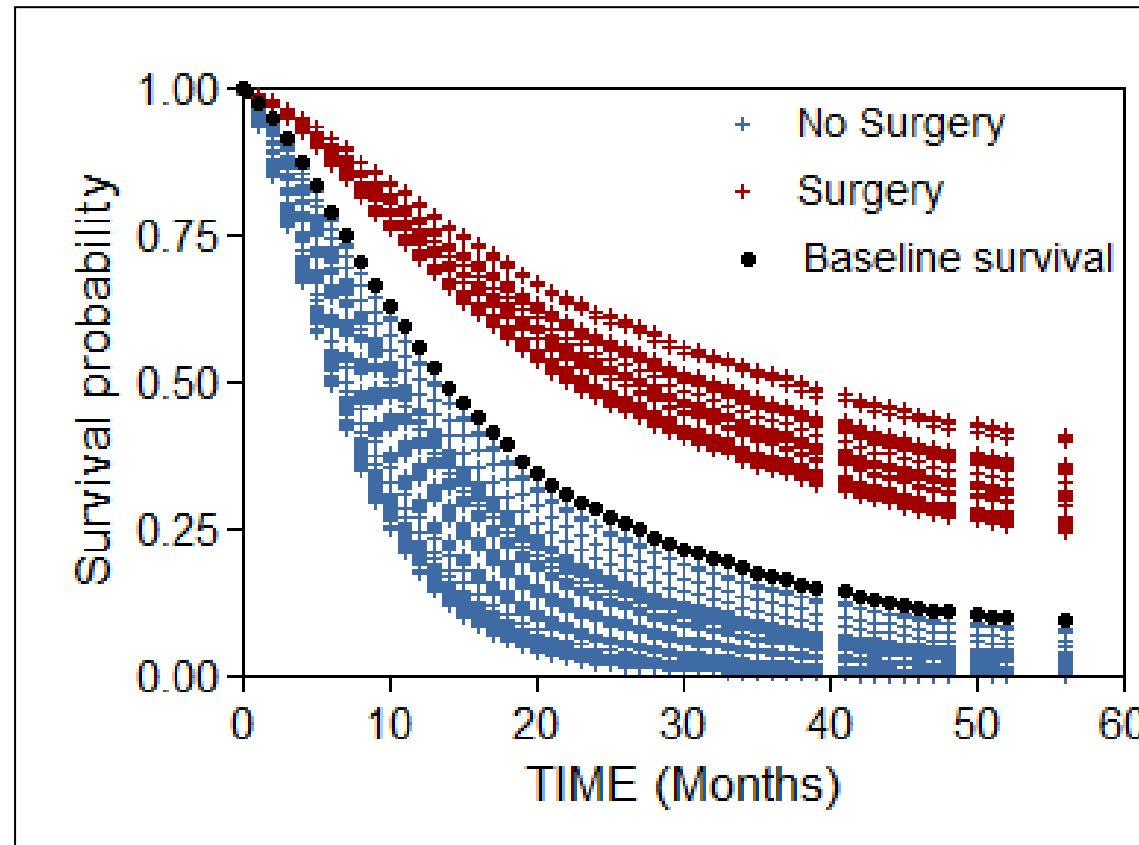
Figure 5. Diagnostic plots for evaluation of parametric model. A) Cox-snell residual for overall fit of gamma distribution model for SEER-2 data, B) Model fit was assessed using deviance residuals with time for SEER-2, C) Cox-snell residual for overall fit of log-logistic distribution for SEER-1. Cox-snell residual shows that the model fit for SEER-2 was better with gamma distribution. Survival time in SEER-1 data however is best explained by log-logistic distribution.

References

- <http://seer.cancer.gov/> - National Cancer Institute SEER database
- Welch, S. A. and M. J. Moore (2007). J Clin Oncol **25**(16): 2159-2161.
- American Cancer Society (2012).: Cancer Facts and Figures 2012. Atlanta, GA
- Hosmer, D. W. and S. Lemeshow (1999). Applied survival analysis : regression modeling of time to event data. New York, Wiley.
- Klein, J. P. and M. L. Moeschberger (1997). Survival analysis : techniques for censored and truncated data. New York, Springer.

Prediction of survival probabilities

Predictions of survival probability in patients from Ext-2 dataset using semi-parametric model:



Median survival time for predicted survival range from 7-15 months in patients without surgery and more than 25 months in patients with surgery.

Predictions of survival probability in patients from SEER dataset using parametric model:

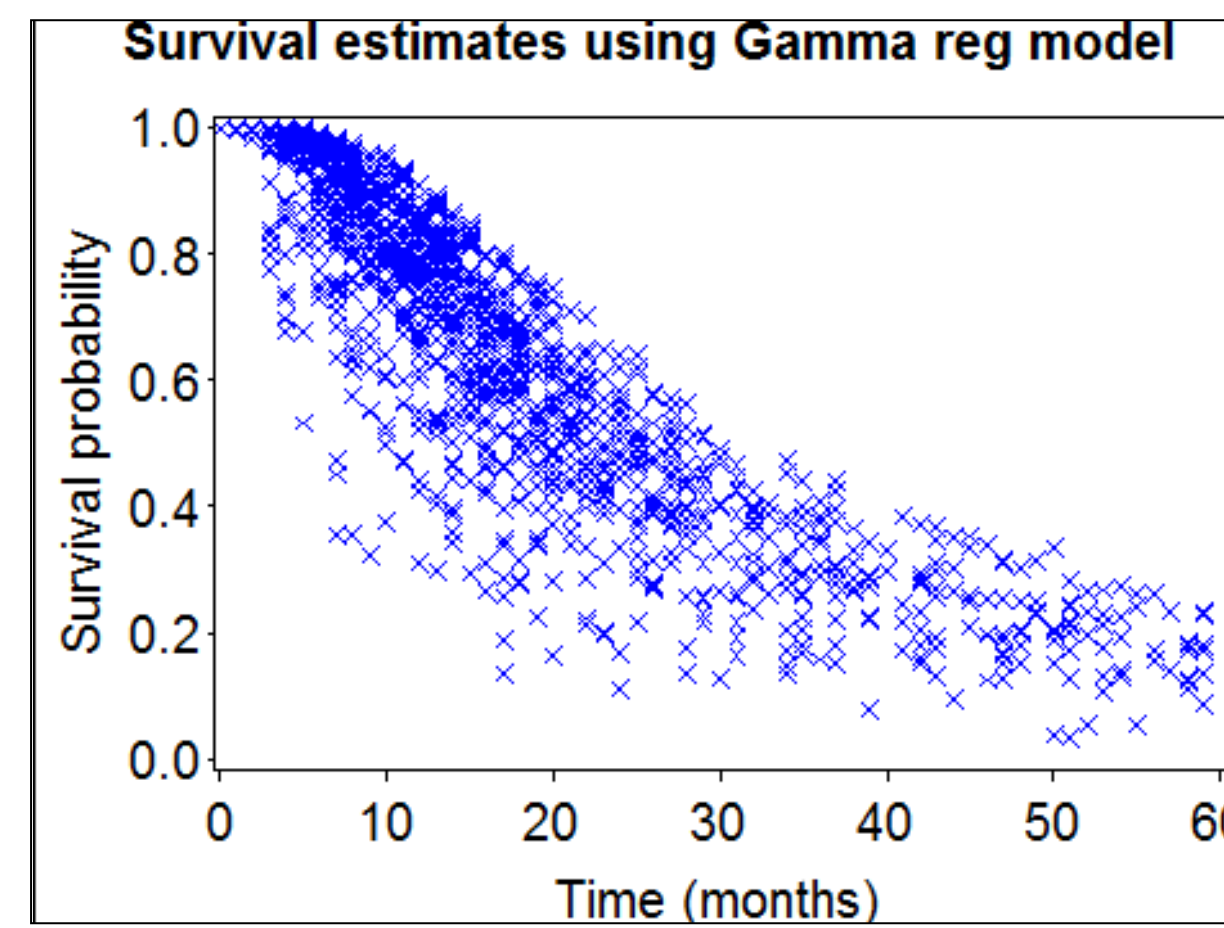


Figure 7. Predictions of survival estimate using parametric model. Figure shows the predicted survival estimates of subjects in the SEER data using parametric survival model.

- Parametric model can also be used to predict the probability of survival in each subject with specific covariates at particular time point (data not shown). This information can be helpful to stratify patients for future trials.
- The median survival time for predicted survival range from 7 – 30 months based on the covariates. Future studies will include cross-validation of the dataset.

Conclusion

- SEER dataset lacks granularity in terms of time dependent information, but provides valuable information on disease state (tumor size, grade, stage, lymph node status) and therapy (surgery and radiotherapy).
- Information on chemotherapy is not complete, however, we used this information with an assumption that subject received standard treatment. Model can be further refined by selecting patients after a particular year when a known treatment was used as standard.
- Median overall survival in PAC patients with non-resectable tumor and chemotherapy is 5 months which is close to previous studies and external dataset.
- Important predictors that influence survival were tumor grade, stage, size, lymph node status, surgery, radiotherapy and chemotherapy.
- Semi parametric analysis of SEER data shows a good fit and the model was used to predict the survival from Ext-2 dataset. Median survival time of patients from predictions were similar to observed median survival time from Kaplan-Meier analyses.
- Parametric analysis of SEER-2 dataset follows gamma distribution. Prediction of survival probability at a particular time can be used to for selecting patients under clinical trials.

Acknowledgements

We would like to thank Piramal Life Sciences for funding postdoctoral fellowship. We also thank Drs. Banke Agarwal (St. Louis University) and Andrew Ko (University of California, San Francisco) for providing the clinical dataset of PAC, and helping with the selection criteria of PC patients in SEER dataset.