23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 1 to 6 December 2019
mssanz.org.au/modsim2019

A flexible family of hypertabastic models

K. P. Singh[a], H. N. Ndetan[a], Z. Bursac[b], W. M. Eby[c], M. A. Tabatabai[d], M. Eum[e], A. A. Bertolucci[f] and S. Bae[g]

[a] Department of Epidemiology and Biostatistics, The University of Texas Health Science Center at Tyler, Tyler, TX 75708, USA, [b] Department of Biostatistics, Florida International University, Miami, FL 33199, USA, [c] Department of Mathematics, New Jersey City University, Jersey City, NJ 07305, USA, [d] School of Graduate Studies and Research, Meharry Medical College, Nashville, TN 38163, USA, [e] Departments of Nursing, Daewon University College, Jaechun, Chungbuk, R.O.K, [f] Department of Biostatistics, The University of Alabama at Birmingham, Birmingham, AL 35294, USA, [g]Department of Medicine, The University of Alabama at Birmingham, Birmingham, AL 35294, USA
Email: Sbae@uabmc.edu

Abstract: Survival analysis is a major tool in cancer research, with a wide application in modeling a variety of cancer survival time data. The family of hypertabastic models includes the hypertabastic proportional hazards model and the hypertabastic accelerated failure model. The hypertabastic survival model has been applied to analysis of various types of cancer data including breast cancer, multiple myeloma, and glioma and to the analysis of non-cancer data. In the area of medical genomics, Tabatabai et al analyzed breast cancer data using clinical and multiple gene expression variables using the hypertabastic proportional hazards model and compared the results with Cox regression. Compared with Cox regression, the increase in accuracy was complemented by the capacity to analyze the time course of disease progression using the explicitly described hazard and survival functions. Recently the hypertabastic accelerated failure models have also been used to analyze mylar-polyurethane insulation data. This gives a new dimension in the application of hypertabastic survival models in biomedical settings. In his paper, we discuss the family of flexible hypertabastic models with applications in cancer.

Keywords: *Time-to-event data, proportional hazards model, hyperblastic models, goodness of fit test, multiple gene expression*

1. INTRODUCTION

Survival analysis is a major tool in biomedical research, as it has a wide application in modeling of all types of cancer survival times. The Cox proportional hazards (PH) model (Cox 1972) is a commonly used model for analysis of survival data. However, the assumption of proportionality of hazards for all covariates in the Cox PH model is needed, though it is not often checked, Atman et al. (1995). In the Cox model, semiparametric in nature, the baseline hazard function is regarded as a nuisance parameter, while in parametric models, the hazard function reflects the time course of the process under study. Estimation of the hazard function is useful in the analysis of change-point hazard rate models. It is known that parametric survival estimates may be more precise than Kaplan-Meier estimates when there are few patients in a particular stratum. Also, parametric or semi-parametric models in survival analysis lead to smoother and more accurate estimators of the hazard and survival functions.

Commonly used parametric time-to-event models or distributions are the Weibull, log-logistic, and log-normal distributions. The accelerated failure time (AFT) models consist of the log-logistic and log-normal distributions for modeling non-monotone hazard rates, Lawless (2002). However, the Weibull model is not appropriate when the hazard rate is non-monotonic. Due to the symmetric property of the log-logistic model, the model may be poor for the cases where the hazard rate is skewed or heavily tailed. The mathematical simplicity of the log-normal model for survival data, especially with right censored observations, is not attractive, Bennett (1983). The data can be better explained by examining the parameter values of the best fitted model.

Tabatabai, et al. (2007, 2012A, 2012B) proposed hypertabastic survival distributions which include the hypertabastic proportional hazards model and the hypertabastic accelerated failure time (AFT) model. Hypertabastic models are great alternative tools in the analysis of time-to-event data in biomedical and other sciences. In this paper, we review the family of the hypertabastic models and discuss the flexibility of these models in modeling survival, or in general, time-to-event data. The use of the hypertabastic models for survival analysis provides additional tools and methods beyond those available through Cox regression. In addition to the increased accuracy provided by the hypertabastic model, it is also possible to give explicit functions describing the time course of both hazard and survival. The explicit survival functions can be used to compute probabilities of survival for a given time for a patient with any profile given the relevant covariates.

2. A FAMILY OF HYPERTABASTIC SURVIVAL MODELS

For modeling time-to-event data, Tabatabai et al. (2007) proposed a two-parameter hypertabastic distribution. Tran (2014) and Tahir et al. (2017) considered hypertabastic models for modeling survival data on several cancer related applications.

2.1 The Hypertabastic Proportional Hazard Model

A continuous random variable t has a hypertabastic distribution if its cumulative distribution function is

$$\text{defined by } F(t)=\begin{cases} 1-sech\left\{\alpha\left[1-t^{\beta}coth\left(t^{\beta}\right)\right]/\beta\right\} & t>0 \\ 0 & t\le 0. \end{cases} \tag{1}$$

And the hypertabastic probability density function is given by

$$f(t)=\begin{cases} sech[W(t)]\left[\alpha t^{2\beta-1}csch^{2}(t^{\beta})-\alpha t^{\beta-1}coth\left(t^{\beta}\right)\right]tanh[W(t)] & t>0 \\ 0 & t<0 \end{cases} \quad \text{where } W(t)=\alpha\left[1-t^{\beta}coth(t^{\beta})\right]/\beta. \tag{2}$$

Furthermore, the Hypertabastic Proportional Hazard Model takes the form of

$h(t\mid x,\theta)=h_0(t)g(x\mid\theta)$ where $h_0(t)$ is the baseline hazard function, given by

$$h_0(t)=\alpha\left[t^{2\beta-1}csch^{2}(t^{\beta})-t^{\beta-1}coth(t^{\beta})\right]tanh\left[W(t)\right] \text{ and } g(x\mid\theta)=Exp\left[\sum_{k=1}^{p}\theta_k x_k\right]. \tag{3}$$

The Hypertabastic Survival Function, $S(t \mid x, \theta)$, for the proportional hazards model has the form

$S(t \mid x, \theta) = \left[S_0(t)\right]^{g(x|\theta)}$ where $S_0(t)$ is the baseline survival function, given by $S_0(t) = sech\left\{\alpha\left[1 - t^{\beta} coth(t^{\beta})\right]/\beta\right\}$. (4)

From the characteristics of the hazard rate function, the function can be monotonic. It may be increasing (I), decreasing (D), or ∩-shape. Tabatatabai et al. (2007) state that the hazard characteristics of the hypertabastic hazard function are as follows:

a) If $0 < \beta \leq 0.25$, the hazed rate ↓from ∞ to 0. b) If $0.25 < \beta \leq 1.00$, the hazed rate is unimodal (∩-shape).

c) If $1.00 < \beta \leq 2.00$, the hazed rate ↑with upward concavity until reaching the inflection point, then continues ↑ with downward concavity. d) If $\beta > 2.00$, hazard rate ↑ with upward concavity.

One of the unique strengths of the hypertabastic model is the flexibility of the hazard function that permits the data to determine the nature of the hazard function without its being inadvertently imposed through the selection of an improper model. The consideration of the different hazard shapes explains the different biological mechanisms of the disease progression. This helps clinicians and researchers to understand the disease status over time.

2.2 The Hypertabastic Accelerated Failure Time Model

The hypertabastic distribution can be used to analyze the accelerated hazards regression model of Chen (2001). The model is called the accelerated failure time model when the covariates interact multiplicatively on the time-scale, Kleinbaum and Klein (2005) and Collet (1994). A hazard function $h\left(t\middle|X,\theta\right)$ in the form, $h\left(t\middle|X,\theta\right) = h_0\left(tg\left(X\middle|\theta\right)\right)g\left(X\middle|\theta\right)$, is assumed in the hypertabastic accelerated failure time model, where $h_0(\bullet)$ is the baseline hypertabastic hazard function. The hypertabastic survival function for the accelerated failure time model is given by $S\left(t\middle|X,\theta\right) = S_0\left(tg\left(X\middle|\theta\right)\right)$ where $S_0(\bullet)$ is the baseline hypertabastic survival function. Finally, the hypertabastic probability density function for the accelerated failure time model is

$f\left(t\middle|X,\theta\right) = f_0\left(tg\left(X\middle|\theta\right)\right)g\left(X\middle|\theta\right)$ where $f_0(\bullet)$ is the baseline hypertabastic probability density function.

The maximum likelihood function method may be used to estimate the parameters in this model. Commonly, though not exclusively, time-to-event data in survival analysis employs right censoring. Detailed estimation of the parameters can be found in Tabatabai, et al. (2007, 2012A, 2012B)

3. SIMULATION STUDIES

Tabatabai et al. (2007) evaluated the performance of the hypertabastic model by conducting a simulation study and compared the overall fit of the proposed model with Weibull, log-logistic and log-normal models. Since all distributions under consideration had exactly two parameters, the negative of the log-likelihood was used as a measure of goodness-of-fit. This measure would result in the same conclusion as the Akaike's Information Criterion (AIC), Akaike (1974). The authors conducted 1000 simulations with a sample size of 200 and random censoring of approximately 40%. A time-to-event data set was generated from 11 different parameter combinations of two parameter Weibull, log-normal and gamma distributions for a total of 33 combinations or 33,000 simulations. The four models were fitted and the -log likelihood averaged over 1000 runs to determine which model best fitted the simulated data with the overall most precision on the average. In simulations, samples were generated from a two-parameter Weibull distribution; the Weibull model fit the data with highest precision in all instances. The hypertabastic model was a close second, outperforming log-normal and log-logistic models for all combinations of parameters, with the log-normal being the worst.

Similarly, when sampling was from a two-parameter log-normal distribution, the log-normal model outperformed all other models. The hypertabastic model and the log-logistic showed similar results, with the log-logistic being slightly better in eight combinations of parameters and the Weibull model performing the worst. Finally, when sampling from a two parameter gamma distribution, the Weibull model fit with the most precision in seven out of eleven combinations. That is because the Weibull distribution and the gamma distribution have hazard functions which are similar in shapes. In another instance, the hypertabastic model slightly outperformed the Weibull model. The log-logistic model came in third; however, it was close to the hypertabastic and the Weibull for several combinations, while the log-normal did the worst in all instances.

Simulation studies with this model in Bursac et al. (2007) demonstrated some degree of robustness with respect to variations in the distribution of the data. Simulations have shown it to be robust with respect to departure from distribution. The feature of the hypertabastic distribution in adjusting its shape is a more accurate representation of the time course of the hazard and survival functions. In the context of the current work of scientists in developing gene expression variables for clinical use, these novel features of this model become even more significant, Bursac et al. (2009).

Tran (2014) and Tahir et al. (2017) proposed a generalized chi-squared test statistics for complete, censored and censored with covariates data in survival analysis. The authors considered the flexible parametric models, evaluating their statistical significance by using their proposed test and log-likelihood test statistics. These parametric models include the AFT and PH models based on the hypertabastic distribution. The authors designed simulation studies to demonstrate the asymptotically distributed normal distribution of the maximum likelihood parameter estimators of hypertabastic model, and validated the asymptotic property of their generalized test statistic for the hypertabastic distribution when the right censoring probability equals 0% and 20%.

4. APPLICATIONS

4.1. Analysis of Breast Cancer Clinical and Gene Expression Interaction Data

Recent focus of research is using gene expression as a predictor of outcome in cancer patients, and improving prognostic capabilities using genomic information. For analysis of gene expression data, the semi-parametric Cox proportional hazard model and the Kaplan-Meier estimator for the survival and hazard curves are utilized. Tabatabai at el. (2012B) applied hypertabastic survival models to the 295 patients from the Netherlands Cancer Institute which is presented in van de Vijver et al. (2002) as a validation set for the seventy gene signature. In selecting from among the hypertabastic, log-logistic, and Weibull proportional hazard models, they compared these models using the −2 log-likelihood score and the Akaike Information Criterion (AIC).

The authors used the following variables in their analyses: Clinical variables-estrogen receptor status (ERS), tumor grade (TG1 and TG2), age (AGE), tumor diameter (DIAM), and lymph node status (LN1 and LN2). The primary tested gene expression variable was the seventy gene signature (70G) of Van't Veer et al. (Nature 2002), which selected genes for prediction of early distant metastasis. From the study of the wound healing microenvironment by Chang et al. (PLoS Biology 2004, *PNAS* 2005), the wound response signature (WRS) and the core serum response correlation (CSR) were included as potential gene expression variables. The core serum response was developed in Change et al. (2004) to represent a canonical expression of fibroblasts activated by serum, and it is a cell-cycle independent set of genes in areas including vascularization, cell motility, and matrix remodeling, common to both the wound healing and tumor microenvironments. Finally, in the area of gene expression for classification of molecular subtype, the authors considered correlation used for validation in van de Vijver, et al. (2002) (CVal), and with centroids for normal (CNorm), ErbB2+ (CERBB), Lumina A (CLumA), Lumina B (CLumB), and basal (CBas) from Sotiriou et al. (2003).

Hypertabastic survival models provided the best fit among all the models considered. Use of multiple gene expression variables also provided a considerable improvement in the goodness of fit of the model, as compared to use of only one. By utilizing the explicit survival and hazard functions provided by the models, the magnitude of the maximum rate of increase in hazard, and the maximum rate of decrease in survival, as well as the times when these occurred were determined. The influence of each gene expression variable on these extrema was explored. Furthermore, in the cases of continuous gene expression variables, represented by a measure of correlation, Tabatabai at el. (2012B) were able to investigate the dynamics with respect to changes in gene expression.

By using parametric hypertabastic survival models proposed by Tabatabai et al. (2007), Tabatabai et al. (2012B) showed the advantages that can be gained by utilizing parametric models, which allows use of explicitly defined, continuous hazard and survival functions for tools in analysis. Using the explicit hazard and survival functions provided by these models, the authors demonstrated some of the potential for analysis of temporal dynamics of the progression of hazard and decrease in survival. The survival function can be used to explicitly compute probability of survival to a given time, and this prediction takes into account an individual patient's profile with respect to any significant variables included in the model.

In summary, the use of three different gene signatures in the model provided a greater combined effect and allowed the authors to assess the relative importance of each in determination of the outcome in this data set. These results point to the potential to combine gene signatures to a greater effect in cases where each gene signature represents some distinct aspect of the cancer biology. The hypertabastic survival models can be an effective survival analysis tool for breast cancer patients.

5. CONCLUSIONS

In this paper, we discuss a family of hypertabastic survival models which include hypertabastic proportional hazards models with a parametric baseline hazard function, and the hypertabastic accelerated failure time models. The hypertabastic hazard function can assume shapes. It can be used to analyze biomedical data such as cancer recurrence time. It can be used to monitor disease progression and regression and provide clinicians with the time interval(s) in which the disease progresses or regresses and in which progression or regression speeds up or slows down. This vital information will make it easier for physicians to take appropriate action regarding their patients. The applications of the family detailed in this paper illustrate the usefulness of the family by modeling various types of cancer data, including analysis of the survival of breast cancer patients, exploring the role of a metastasis variable in combination with clinical and gene expression variables.

The family is flexible in shape and robust to various underlying distributions. We recommend that clinicians, practitioners and data analysts consider comparing this model to other common survival models, prior to deciding which one provides the best fit and prediction. The use of the family of hypertabastic models for survival analysis provides additional tools and methods beyond those available through Cox regression. In addition to the increased accuracy provided by the hypertabastic models, it is also possible to give explicit functions describing the time course of both hazard and survival. The explicit survival functions can be used to compute probabilities of survival for a given time for a patient with any profile given the relevant covariates. The explicit survival and hazard functions determined from the hypertabastic models allow the analysis of the time course of both of these functions and their graphical representation. Furthermore, the explicit survival functions allow for computation of survival at any given time for a patient with any specific covariate profile. In conclusion, the authors have presented a new family of innovative models of survival data analysis, which is an important aspect in data analysis.

REFERENCES

Akaike H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*. 19(6): 716-723. doi:10.1109/TAC.1974.1100705.

Bursac Z., Tabatabai M. A., Williams D. K., Singh K. P. (2009). Simulation Study of Performance of Hypertabastic Survival Models in Comparison with Classic Survival models. *The 2008 ASA Proceedings of the Joint Statistical Meetings, Biometrics Section,* pp. 617-22. Alexandria, VA.

Chang H. Y., Sneddon J. B., Alizadeh A. A., et al. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biology*; 2:206–214.

Chen Y. Q. (2001). Accelerated hazards regression model and its adequacy for censored survival data. *Biometrics*. 57:853-860.

Collet D. (1994). Modeling Survival Data in Medical Research. *Chapman and Hall/CRC.* New York.

Cox D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society*. Series B, 34:187-220.

Kleinbaum D. G., Klein M. (2005). Survival Analysis. A Self-learning Text 2nd edition. *New York, Springer*.

Lawless J. F. (2003). Statistical models and methods for lifetime data (Second edition). *Wiley, New York.*

Tabatabai M. A., Bursac Z., Williams D. K., Singh K. P. (2007). Hypertabastic survival model. *Theoretical Biology and Medical Modelling*. 4:40. doi:10.1186/1742-4682-4-40.

Tabatabai M. A., Eby W. M., Nimeh N., Li H., Singh K. P. (2012). Clinical and multiple gene expression variables in survival analysis of breast cancer: Analysis with the hypertabastic survival model. *BMC Medical Genomics*. 5:63 doi: 10.1186/1755-8794-5-63.

Tabatabai M. A., Eby W. M., Nimeh N., Singh K. P. (2012). Role of metastasis in hypertabastic survival analysis of breast cancer: Interaction with clinical and gene expression variables. *Cancer Growth and Metastasis*. 5:1–17. doi: 10.4137/CGM.S8821.

Tabatabai M. A., Bae S., Singh K. P. (2015). Analysis of survival data using hypertabastic models. *The Proceedings of the International Statistical Institute World Congress*. pp. 4007-12.

Tahir M. R., Tran Q. X., Nikulin M. S. (2017). Comparison of hypertabastic survival model with other unimodal hazard rate functions using a goodness-of-fit test. *Statistics in Medicine*. DOI: 10.1002/sim.7244.

Tran Q. X. (2014). Dynamic regression models and their applications in survival and reliability analysis. PhD Thesis. University of Bordeaux, France.

Van de Vijver M. J., He Y. D., van't Veer L. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *NEJM*. 347.1999–2009.

Van't Veer LJ, Dai H, van de Vijver MJ, et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 415:530–536.