Received: 2 October 2019 | Revised: 14 January 2020 | Accepted: 3 February 2020

DOI: 10.1002/wics.1504

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Review of current advances in survival analysis and frailty models

Usha S. Govindarajulu[1] | **Ralph B. D'Agostino Sr.**[2]

[1]Center for Biostatistics, Department of Population Health, Icahn School of Medicine at Mount Sinai, New York, New York

[2]Department of Mathematics and Statistics, Boston University, Boston, Massachusetts

Correspondence
Usha S. Govindarajulu, Center for Biostatistics, Department of Population Health, Icahn School of Medicine at Mount Sinai, New York, NY.
Email: usha.govindarajulu@mountsinai.org

Abstract

In this article, we have presented a review of existing methods and trends in survival analysis and frailty models. The background has been presented for each topic discussed for survival and frailty models where the presentation flows from original methods to more advanced methods. This article has also shown various current methodologies that exist among survival and frailty models. The advantages and disadvantages of more recent methodologies are presented and discussed in this review.

This article is categorized under:
Statistical Models > Survival Models
Statistical Models > Semiparametric Models

KEYWORDS

competing risks, copulas, Cox model, cure models, frailty models, recurrent events, survival analysis

1 | INTRODUCTION

Survival methods have traditionally been designed to handle time-to-event analyses where the outcome is time to the event of interest. Frailty models were adapted from survival models to add in a random effect term to account for unexplained heterogeneity. Both survival models and frailty models have received considerable attention in the literature and have been adapted and changed over the years as the field has moved forward. We have discussed and reviewed some of the trends in these models, especially in the many adaptations to the Cox proportional hazards (PH) regression model and the incorporation of frailty with newer forms of modeling like recurrent events models, copula models, and cure models.

2 | METHODS

2.1 | Traditional survival models

Survival analysis has been a field in which time to event analyses are conducted. The outcome has been unique to this field where time becomes the outcome. The main terms denoted in this field are the hazard rate, $h(t)$ and survival rate, $S(t)$, and these can be manipulated through various equations with the probability density function, $f(t)$, the cumulative hazard function, $H(t)$, and the cumulative distribution function, $F(t)$. The $S(t)$ was specified as a probability or a proportion of persons surviving up until time T and can be denoted as $S(t) = P(T > t) = 1 - F(t)$. Meanwhile, the hazard rate,

$h(t)$, was not a probability but rather an instantaneous risk to have an event given one is event free up until that time (Hosmer, et al., 2008):

$$h(t) = \lim_{\Delta t \to 0^+} \left\{ \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} \right\} = \lim_{\Delta t \to 0^+} \left\{ \frac{P(t \le T < t + \Delta t)}{\Delta t P(T \ge t)} \right\} = \frac{f(t)}{S(t)}. \quad (1)$$

Equation (1) also has demonstrated how $h(t)$ can be converted between $f(t)$ and $S(t)$. For $H(t)$, this can be obtained by integrating $h(t)$ from 0 to t and this has been why $H(t)$ is often referred to as the integrated hazard rate.

In addition, these models have incorporated censoring of subjects from an analysis where an observation can be right censored from an analysis if the subject has the event of interest after the end of study time or some other competing risk occurs to cause the subject to drop out of the study. Competing risks are discussed later in this review. There have been four types of possible censoring: left, right (the most common), double, and interval. Censoring was denoted by occurrence: prior to the study (left), by the end of study (right), as a combination of both left and right (double), or during defined sequences of time (interval). In addition, subjects could have been truncated from analyses where they could be truncated out even before the investigator is aware of their existence (left) or if they never have the event of interest during the period of follow-up (right). Right censoring has been the most common form of censoring and when there is a combination of right censoring, which is fixed, and censoring due to competing events, than it is referred to as random right censoring.

An important assumption in survival analysis which has often been forgotten and over-looked has been the non-informative assumption, which implies that censoring should be independent of the true event time, though these two assumptions may not be equivalent (Kleinbaum & Klein, 2011). However, this has been a critical assumption of the majority of the survival methods, yet, this assumption has been significantly under-appreciated.

In general, one of the main uses of survival analysis has been the nonparametric methods introduced decades ago by Kaplan–Meier (Kaplan & Meier, 1958), where $S(t)$ is calculated at each unique event time by subtracting number of events from number still at risk and then dividing by number still at risk. In this way, it has functioned as a counting process. It has been common practice to graphically view the Kaplan–Meier survival curves plotted against time for which we show an example of a Kaplan–Meier survival plot, which is always plotted as a step function. A drop would have occurred at each unique event time when there is an event and hatch marks have denoted censored observations, which we have shown in an illustrative example (Figure 1). From this method, survival analysis expanded to the Log Rank test, which finally allowed comparing Kaplan–Meier survival curves statistically. This was first developed by Mantel (1966) and named by Peto. However, these were unadjusted survival estimates and comparisons and did not allow for adjustment by possible confounders or effect modification.

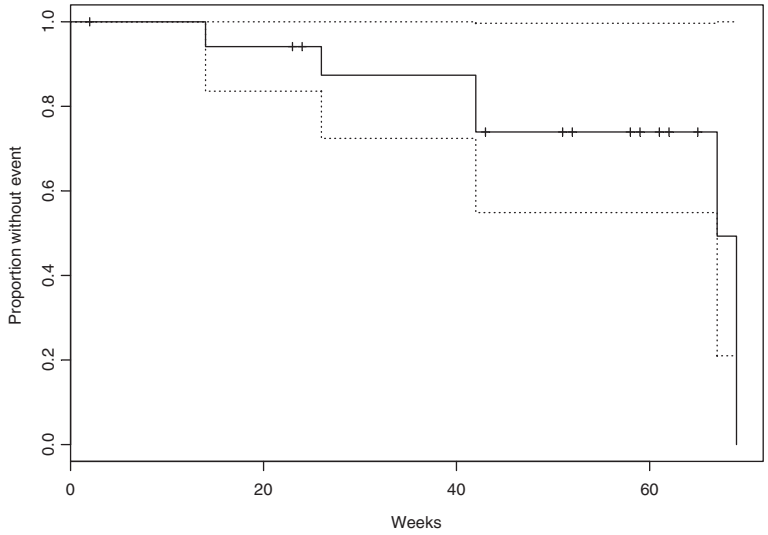

FIGURE 1 Kaplan–Meier survival curve example

The field then expanded into parametric regression to be able to adjust for confounders and effect modification through regression. The main caveat with parametric regression was specifying the distribution of the survival times and therefore, the user needed to have some idea of what appropriate distribution would allow for positive real values to use to model the survival times. Though there were many options besides the exponential distribution, which only allows for a flat baseline hazard rate some of these other options for parametric distributions that have existed are the Weibull, log-normal, Gompertz, log-logistic, and so forth. The Weibull and exponential distributions required the assumption of PH, a property which will be discussed in more detail with the next method. The Weibull at least allowed for monotonically increasing or decreasing survival rates so was an improvement to the exponential hazard, which was flat. Other distributions like the log-normal and log-logistic at least relaxed this assumption and allowed for nonmonotonically increasing or decreasing hazard curves and therefore, more richness of shapes of the survival curve. These models were often represented in an accelerated time format (AFT) which allowed describing the survivor relationship between any two observations and where the natural logarithm of time was the outcome. The coefficients in an AFT model could also be converted in order to describe the hazard rate for coefficients. In terms of the next topic though, since modeling the hazard rate became more popular, we have not discussed as much the AFT format for more advanced applications.

The advent of the Cox PH regression model (Cox, 1972) was quite revolutionary in the survival analysis world. It also brought up a new term that had not been used before and for the most part, has been unique to this model to this day, semiparametric. This term highlighted the advantages of the method having both nonparametric and parametric properties. Even today, this method is widely adopted for most survival analyses. The nonparametric part of the model contained the unspecified baseline hazard rate, $h_0(t)$, which for the parametric models always had some form of a distribution specified. It existed as a sinkhole for the intercept term or, essentially, where all values of the covariates were equal to zero. The parametric portion was modeled similar to regular regression with an additive combination of predictors, each with a corresponding estimated coefficient in the model. The Cox PH regression model for the ith observation has the following equation:

$$h_i(t) = h_0(t)exp(\beta' x_i), \tag{2}$$

where the function of the covariates is multiplied onto the baseline hazard rate to give the overall hazard equation for the model. The model has appeared simplistic in notation but has been very powerful. However, the model came with a major caveat which we next describe.

The major assumption of the model has been the criteria of PH, which is a fairly strict assumption. The assumption was that for two observations, the hazard rate between the observations varies over time per some constant. Per this assumption, this was how Sir David Cox could leave the baseline hazard rate unspecified. Various methods came about to validate this assumption. The easiest way to understand the PH was through plotting $S(t)$. Plots of the $\ln(-\ln(S(t)))$ against time would show the survival curves as linear and allow seeing if the proportional assumption was met. This was one such method in which the PH assumption could be visualized. We have presented an example of these plots, which is generated from data from a randomized clinical trial (Nduati, et al., 2000), conducted November 1992–July 1998 in Nairobi, Kenya on vertical transmission of HIV in either breast feeding infants or formula feeding infants. These plots have shown clear violation of the PH assumption with the crossing curves (Figure 2).

Another method, Schoenfeld residuals (Schoenfeld, 1982), which came about, could be used for testing globally and locally the effect of the PH assumption. It was used for plotting the residuals to test the assumption but also for significance testing of the effects of the PH assumption, which was done overall and for each predictor separately. This helped to determine which term(s) was causing the nonproportionality. However, then the question arose of how this assumption could be handled if it were violated and if the Cox model could at all be used in any fashion. Two different ways of handling this then came about: stratified PH modeling and the use of time-dependent covariates. We have discussed each type separately in the upcoming sections. Before these are explained, we have shown below that there is another format for presentation of survival models called counting process format, though typically survival models are presented in the single observation format.

2.1.1 | Counting process format

Survival modeling has been divided in the world of its presentation. A counting process formulation for survival analysis was noted by Fleming and Harrington (1991) and Andersen, et al. (1993). This notation was motivated by

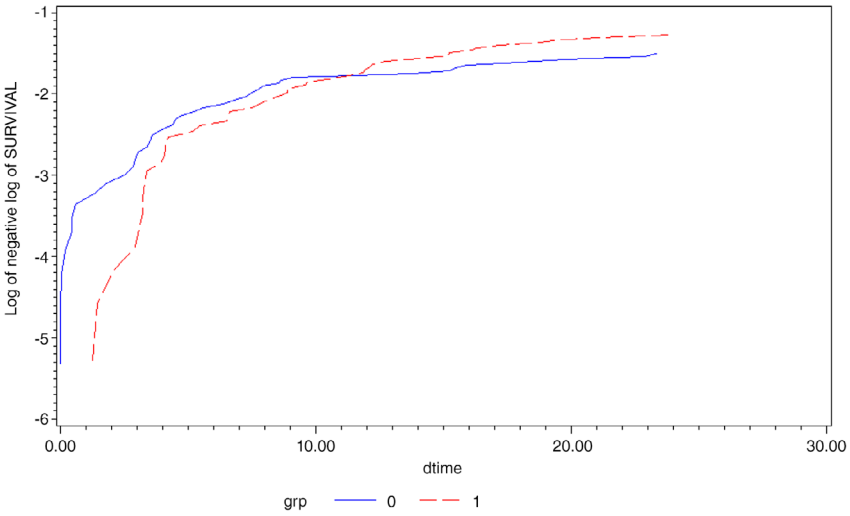

FIGURE 2 −ln(ln(survival)) plots example to test proportional hazards assumption

considering the survival events in times as being accrued through counting processes, similar to how they are calculated in the nonparametric methods like Kaplan–Meier. The notation of the survival process through the counting process format became very useful for adaptation of a person having more than one event and for time-varying covariates to be discussed below. As we will see in the next main section, it also became especially useful for models with repeated observations, and also for those models with random effects. The following equation has been shown as the format of the Cox model in counting process format:

$$d\Delta(t,X_i(t)) = Y_i(t)\exp(\beta' X_i(t))d\Delta_o, \tag{3}$$

where the counting process that exists for each person has been denoted as N_i, which is retained as a counter of the number of observed events for each subject over the follow-up period. The counts were accrued across steps, similar to the way the Kaplan–Meier was handled as a counting process. In this format, instead of the hazard rate which is based on continuous values in a defined time period, a cumulative intensity process is modeled, which is based on accruing rates over time for more than one possible time period and denoted by Λ. Finally, we have shown there is also an indicator, Y_i, for whether or not the subject is at risk.

2.1.2 | Stratified Cox models

The stratified Cox PH regression model has been represented as the same Cox PH regression model but with a stratification occurring on the baseline hazard rate. Kalbfleisch and Prentice (1980) added this unique feature to the Cox model. A predictor of less interest that might be a variable used to control randomization or used for adjustment may be considered to be used as a stratification variable in the model. The prior equation for the Cox model (Equation (3)) has been adapted to account for the stratum, which is denoted as s in the baseline hazard rate in Equation (4):

$$h_{ij}(t) = h_{0s}(t)\exp\left(x_{ij}\beta\right). \tag{4}$$

Controlling for a particular variable that was causing the nonproportionality by making it a stratum variable would handle the nonproportionality aspect at the same time as making the model have PH. The stratum variable then was usually not estimated at all as it was relegated to the baseline hazard rate, but if there was a need to estimate stratum specific survival rates then this was a possibility (Hosmer et al., 2008).

2.1.3 | Time-varying models

Time-varying covariate models have worked in either one of two ways in order to handle nonproportionality and also allow for a covariate which varies over time (Kalbfleisch & Prentice, 1980). These models have either included an interaction between a predictor, usually one that appears to be causing the nonproportionality issue, and time in the model. Incorporation of this interaction term then allowed for nonproportionality and handled it all at the same time. Another way has been to include a binary indicator term as long as there is another variable, which as a secondary measure of survival time, can be included to create that interaction term in the model. Time-varying models have been adapted to the Cox model and parametric survival models and, therefore, allowed for data that is measured longitudinally. Time-varying covariates are handled well through the counting process format.

2.2 | Frailty models

A large addition to the world of survival models was frailty models. These class of models allowed for the addition of a random effect term to the multiplicative function to handle heterogeneity or unmeasured covariates that the model did not already handle through measured covariates. Some history of these models has been that they were motivated from other fields, like demography by Vaupel, et al. (1979) where he developed the term frailty to describe heterogeneity at the individual level but this was expanded to describe heterogeneity amongst groups of individuals or within an individual, making it multivariate. This was further described below.

Many different variations of the frailty model came into place and extensions into more advanced forms of survival models with frailty came into the foray as well. The easiest type of frailty model to understand has been the univariate frailty model, where each observation has a random component and frailty varies from individual to individual. Generally, this type of model was more for theory and rarely used in practice. The more traditional model came about as the shared frailty model, which allowed for having a variable which indicated a grouping of individuals in a unit like a family or sibling group, a unit that allowed for shared characteristics. The frailty was then modeled on this shared group effect. The frailty effect has arisen from any distribution in the positive range such as gamma, the most mathematically tractable form (Vaupel et al., 1979), lognormal, and other positive distributions. This group effect can be dichotomized to be the shared group effect and individual effect in correlated frailty models as described in (Govindarajulu, Lin, Lunetta, and D'Agostino (2011). We represented the frailty in the parametric or semiparametric hazards model as follows although certainly an AFT model could be adapted to incorporate parametric frailty:

$$h_{ij}(t|x,z) = h_0(t)c\left(\boldsymbol{\beta}' X_{ij} + z'_{ij}\mathbf{w}_{ij}\right), \tag{5}$$

where h_{ij} is the hazard function for the jth observation from the ith cluster where the cluster could be an individual or a group, and $\mathbf{w}_{ij}$ was a vector of random effects associated with the covariates vector z_{ij}, for cluster i and observation j, which usually includes an intercept. The random effects accounted for the heterogeneity effects of the z_{ij}. When z_{ij} included only an intercept, the model in Equation (5) became a simple multiplicative frailty model. The $\mathbf{w}_{ij}$ accounted for the correlation among individuals within a cluster. Each component of the frailty then represented a cluster-specific effect with the respective covariate. This type of model allowed inference of cluster-specific random effects, and potentially greater comprehension and interpretation of the heterogeneity in the population. We have shown an example of a graph of survival curves generated separately from Kaplan–Meier, Cox PH regression, and a frailty model, in a paper by Kosorok, et al. (2004), where the Cox model and frailty model estimates were similar while the Kaplan–Meier deviated at times in the tails. They attributed this deviation to sparse data in the tails while the Cox and frailty models use all the data for estimation and so avoid this issue (Figure 3).

As described in Govindarajulu et al. (2011), the frailties across different clusters are assumed to have a distribution; they then account for unexplained heterogeneity at the cluster level. The frailty term, $\mathbf{w}_{ij,}$ would then be univariate or multivariate. A univariate frailty model would contain a single frailty per cluster with $\mathbf{w}_{ij}$ being a scalar and the same across all different j's within the same cluster i, z_{ij} being an intercept of one. Individuals within a cluster shared a common frailty (also termed shared frailty). In a multivariate frailty model, each cluster then would have two or more frailties, and $\mathbf{w}_{ij}$'s are therefore different across different j's even within the same cluster i—the frailties $\mathbf{w}_{ij}$ can come from a multivariate distribution or has more than one term in it (in the case of correlated frailty described below). The

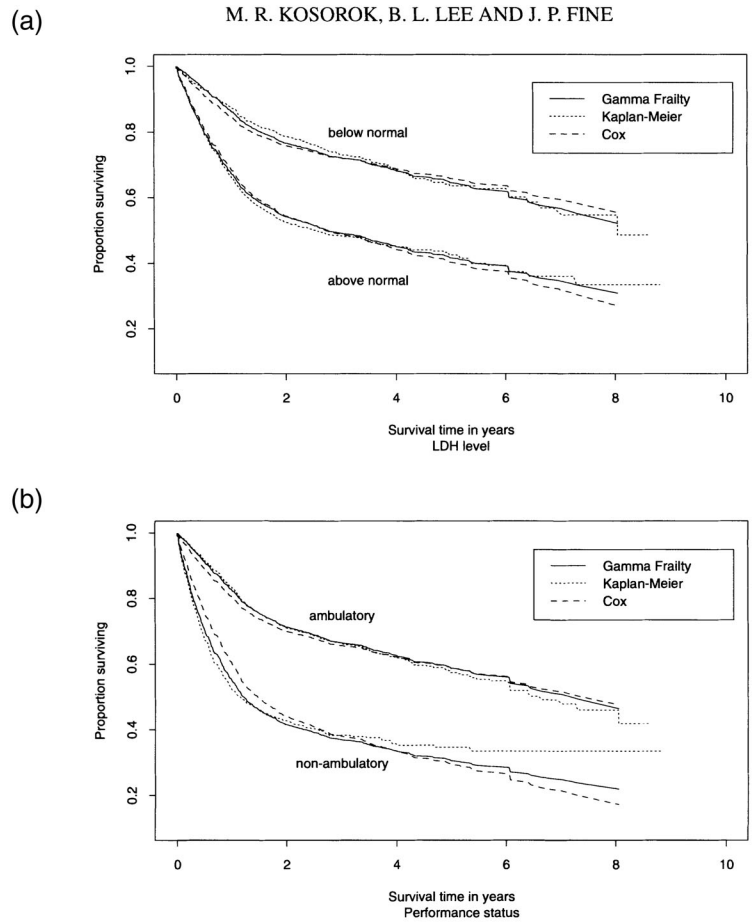

FIG. 1. *Estimated marginal survival distributions of non-Hodgkin's lymphoma patients for LDH level* (a) *and performance status* (b) *under the gamma frailty model. The Kaplan–Meier estimates and Cox model estimates are included for comparison.*

FIGURE 3 Example of survival curves from Kaplan–Meier, Cox model, and frailty model. (a) LDH level. (b) Performance status

frailties among different observations within a single cluster were marginally correlated. In the case of correlated frailty model, two additive terms within $\mathbf{w}_{ij}$ consist of a final $\mathbf{w}_{ij}$. One term was the common frailty term shared among all the observations from the cluster i and the other was the individual frailty not shared; the two frailty terms had two different distributions.

We noted that various methods of estimation of the frailty model exist and are currently used in some software-packages. Some of these methods have been likelihood estimation using: Newton–Raphson, EM algorithm, penalized likelihood, or Bayesian methods (i.e., Monte Carlo Markov Chain), but this list is not exhaustive. These methods essentially have sought to solve the log-likelihood in which the frailty, w, will be integrated out. This has been represented in counting process notation (Govindarajulu et al., 2011):

$$\sum_i \sum_j \left[\int_0^\infty Y_{ij}(t) \left[\log\left(h_0(t) + X_{ij}\beta + Z_{ij}w_{ij}\right) \right] dN_{ij}(t) - \left[\int_0^\infty Y_{ij}(t) \exp\left(X_{ij}\beta + Z_{ij}w_{ij}\right) h_0(t) dt \right] + logp\left(w_{ij};\theta\right) \right], \tag{6}$$

where have an $h_0(t)$ here as an intensity process and not a hazard rate, an indicator, Y_{ij}, for whether or not the subject is at risk, an X_{ij} which contains the matrix of covariates at each event time, and dN_{ij} for sum over number of observed events. The frailty distribution, ρ, is modeled for the frailty, w, with its variance parameter, θ at the end.

The EM algorithm approach for the Cox frailty model (a similar approach exists for parametric frailty) began with the full likelihood being a function of the observed event times and the unobservable frailties. In the two steps of the EM algorithm, the E-step involved computing the expectation of the full likelihood with respect to the observable data

and the M-step invoked a partial likelihood being constructed for estimation of the covariate effects using a profile likelihood technique. One then iterated between the E-step and the M-step until convergence was obtained. A more detailed explanation of the EM algorithm can be obtained from Klein (1992), who also advocates for the counting process formulation for frailty models, or Nielsen, et al. (1992).

Another method of estimation that came about is the use of penalized likelihood estimation. This estimation, which uses penalized regression, has been adapted for estimation of Cox frailty models (Therneau & Grambsch, 2000). The frailty terms were treated as additional regression coefficients which are then constrained by a penalty function added to the log-likelihood. This method was supposedly much faster than the EM algorithm approach. A detailed description was provided by Therneau and Grambsch (2000).

Use of Markov chain Monte Carlo (MCMC) methods in frailty analysis has gained a foothold in the literature as well due to its ability to handle computationally intensive modeling. Chen and Ibrahim (2002) developed a cure model, which will be discussed in more detail later, for multivariate data which incorporates multivariate failure time data for populations with a cure fraction and a frailty term which creates a correlation structure between the failure times. MCMC was then used to sample from the posterior distribution of the parameters. Even for frailty models with doubly censored data, when both initial and final times are censored, Jones (2004) developed a novel way to use MCMC for estimation of this type of model. In modeling the frailty effect as not only a function of unobserved heterogeneity but also observed covariates, Govindarajulu, Glickman, and D'Agostino (2007) also employed MCMC to sample from the posterior distribution of this complex model.

Several others have written extensively on frailty models and published books on these topics. Duchateau and Janssen (2008) have written a comprehensive text on frailty models which also show examples with software. Ha, et al. (2017) have written a detailed text on likelihood inference for correlated survival models.

2.3 | Advanced survival models

2.3.1 | Competing risks models

Competing risk models have allowed for considering more than one event type for each subject. Generally, most survival models have allowed for time to one event; however, these models allow considering a competing event that may occur for a given subject besides the main event of interest. Kalbfleisch and Prentice (2002), in the newer edition of their book, have discussed competing risk models. The competing event framework was adapted into the Cox model. For example, if the main event was death from CVD, a person could experience death from another event other than CVD death. This other event was the competing risk. Since there frequently are competing risks in time-to-event models, these models are being utilized more often in survival modeling to allow for a competing risk or event. Usually, the cumulative incidence function, CIF, is reported from this model, which represents the cumulative proportion of failure at time t due to a specific cause; this then made the cumulative distribution function event specific. This area of research has been rapidly expanding.

2.3.2 | Joint longitudinal and time-to-event models

The motivation from these models arose from having a time-to event analysis along with predictors that are measured longitudinally, which is very common in many prospective studies. For Tsiatis, et al. (1995), the topic was especially motivated by the AIDS epidemic and how to model CD4 counts in a time to progression model. He continued this work with Davidian in further discussing the theory and applications behind this type of modeling, which combines aspects of longitudinal modeling with PH modeling (Tsiatis & Davidian, 2001). Hogan and Laird (1997) had proposed using a mixture model to handle the time-to-event analysis and repeated measures, even in the presence of dropouts. The modeling essentially had encompassed a joint model where a joint likelihood contained the probability density function (pdf) of the time to event variable multiplied by the pdf of the time-varying covariate and the pdf of the random effect. Over time, improvements to the computational complexities of the models have been made. Rizopolus (2012) has discussed the more current state of the mathematical and computational aspects.

2.3.3 | Additive survival models

Traditionally, the parametric survival models and the Cox PH models have utilized the function of the covariates as multiplicative onto the baseline hazard rate. A different type of modeling, additive survival modeling, changed this typical modeling. Instead, the function of covariates has an additive effect to the baseline hazard rate. This was first suggested by Breslow and Day (1987) and then later Aalen (1989) refined the method, making it fully additive and nonparametric, as described by Hosmer et al. (2008):

$$\begin{gathered} h(t,\boldsymbol{x},\boldsymbol{\beta}(t)) = \beta_0(t) + \beta_1(t)x_1 + \beta_2(t)x_2 + \ldots + \beta_p(t)x_p \\ p+1 \text{ fixed covariates and } x' = \left(1, x_1, x_2, \ldots x_p\right). \end{gathered} \tag{7}$$

In this model, rather than viewing individual coefficient estimates, a cumulative regression coefficient was estimated for each covariate instead. These cumulative regression coefficients are then what are also reported for each covariate. Additive modeling has gained a lot of traction in the field and at times is suggested to replace the typical multiplicative hazards framework commonly seen in survival regression models as the additive effect seems to have more biological plausibility.

2.3.4 | Recurrent events models

Frailty models have also been extended into the recurrent events modeling area. This type of modeling has evolved for data where repeated or recurrent events occur within a survival analysis framework. It has almost been the repeated measures analog in the time to event modeling framework. With recurrent events, the correlation between events has to be handled in the estimation of the standard error estimates. A previously discussed format for survival analysis, counting process format, as described by Andersen et al. (1993), has been very useful for representing recurrent events models due to each observation having repeated events.

Prentice, Williams, and Peterson (1981) discussed how to model recurrent events through a stratified Cox model, where the stratum variable was used to keep track of events for each subject. Lin and Wei (1989) showed strong theoretical developments in modeling and especially through their presentation of a robust variance estimator for recurrent events. Staniswalis, et al. (1997) had discussed recurrent events analysis in an applied setting, where they carefully defined recurrent events, and also extended the methodology of Thall's parametric model to include recurrent events. Cook and Lawless (2002) further considered such models by further adapting the frailty PH model to allow for recurrent events. They adapted the Monte Carlo EM algorithm with Metropolis-Hastings sampler to handle the recurrent events and largely expanded on this work. Liu, et al. (2004) have also further extended recurrent events shared frailty models in their work to attend to terminal processes.

2.3.5 | Copulas

Copulas were developed as a way of modeling and have garnered more attention in the survival literature, in general, as well as in the frailty model literature. In probability theory, copulas were multivariate cumulative distribution functions whose marginal distribution for variables is uniform. Therefore, copulas themselves were not unique to survival modeling but they appeared to provide a framework that can be used for multivariate survival models. Georges, et al. (2001) have provided a well-summarized description of the history and the modeling of copulas in their manuscript. They described how Clayton (1978) discussed fairly early on in the literature how to allow for a bivariate association model. In effect, a copula can allow for a joint function of marginal survival models. Clayton defined the copula as where if we let $\check{C}$ be a copula, then we can define a multivariate survival function as follows:

$$\begin{gathered} S(t_1,\ldots,t_N) = \check{C}\left(S_1(t_1),\ldots,S_N(t_N)\right) \\ \check{C}(u_1 \ldots u_d) = \varphi\left(\varphi^{-1}(u_1) + \ldots + \varphi^{-1}(u_d)\right) \\ \check{C}(u) = \left(u_1^{-\theta} + \ldots u_d^{-\theta} - d + 1\right)^{-1/\theta}. \end{gathered} \tag{8}$$

In terms of frailty, the random effect term for the frailty could be incorporated into the copula structure. The first equation (Equation (8)) can be modified in the frailty form where $\check{C}$ could specifically be an Archimedean copula with a generator that conforms to the Laplace transform of the frailty variable. The generator was a function related to the Archimedean copula (Georges, et al., 2001). For Cox PH regression, survival copulas are Archimedean. To describe this a little further, we have shown in the second equation in Equation (8) how the survival copula of T is Archimedean with generator, Φ and in the third equation, which is derived using the second equation, the resulting survival copula is shown if a gamma distribution is used for the frailty random variable.

On the other hand, copulas and frailty models have been contrasted and compared (Wienke, 2010). Goethals, et al. (2008) have noted their similarities but also their differences. They have noted that in the bivariate situation, the copula functions used are the same as compared to a shared frailty model but the marginal survival models are different. Certainly, this makes sense in terms of the fact that the frailty models have a random effect term, which ends up being integrated out of the model and certainly this would alter the marginal survival model structure. However, using a copula model to handle frailty and the dependence of another process has been an area of research with copulas. For example, Emura, Nakatochi, Murotani, and Rondeau (2017) have shown how to have a joint frailty copula model to handle the frailty and also the dependence amongst tumor progression and death. Survival models with copulas and frailty models with copulas have continued to be an expanding field of research.

2.3.6 | Cure models

Cure models have been introduced in survival analysis to handle a fraction of the population who may have been cured and then do not go on and have the event of interest (Sy & Taylor, 2000). This then resulted in a type of mixture of the population making it heterogeneous with those who are cured and those who are not. This mixture has been adapted into the modeling where the fraction being cured is accounted for onto the survival probabilities in the spirit of binomial probabilities.

$$S(t) = \pi^* S(t)_1 + (1-\pi)^* S(t)_2. \tag{9}$$

This implies that there then must have been a mixture of hazards in the model given that there is a mixture of survival probabilities based on the cure fraction. This naturally extended over to frailty modeling, which already handles heterogeneity and which also has a mixture of hazards. Various researchers then started describing cure models combined with frailty as described by Price and Manatunga (2001) who show that the model can handle both the heterogeneity with the frailty distribution but also the cured fraction as well. Wienke (2010), who has done a fair amount of research in copula models and frailty with bivariate survival, has naturally extended the copula framework to cure models. Furthermore, he also extended correlated frailty models to include a cure fraction as well. Cure models have also been adapted to handle recurrent events modeling. Rondeau, et al. (2013) have applied cure models to breast cancer data with recurrences and have been able to adapt the recurrent events modeling with the cure fraction in frailty models. The research with cure models has continued to be a constant growing area of research.

3 | DISCUSSION

In this review article, we have covered two main areas: survival models and frailty models. In survival models, we reviewed the main definitions of survival and hazard, the nonparametric methods like Kaplan–Meier and Log Rank test, and then regression methods with parametric survival regression and Cox PH regression. A lot of expansions that mainly adapt to the Cox model have taken place with time-varying covariates, stratification, recurrent events, competing events, and additive survival modeling. Certainly, all of these techniques are gaining more and more traction in the statistical field and beyond.

Of course, frailty models are part of survival models, but as we have described, they have largely had their own set of developments and have taken off in many developments and applications. Just allowing a random effect term in the model to take on a positive distribution and in effect, handle explained heterogeneity in the model, was a novel development. Oftentimes, developments in survival analysis have been incorporated into the frailty model after they first appeared in survival analysis. Frailty models can be univariate, multivariate, or correlated. The shared frailty model is a

subset of the multivariate frailty and has been modeled quite extensively. Adding to the literature of frailty models, recurrent events were added and handled in the frailty model. In addition, time-varying covariates and competing events have been incorporated with sometimes more than one of these aspects included in a given frailty model. Newer methodological developments in frailty models within the last decade have been seen in joint longitudinal and time-to-event models, joint copula frailty models, and frailty cure models. Combining the frailty with these models has provided to be possible and useful for real data applications.

Unfortunately, the areas of research shown in this article are not exhaustive and we cannot list all the new possible extensions of these modeling that are being formulated at this very given moment, but we have tried to do justice to this ever changing, unique field by providing a broad and at times, detailed overview. New developments in survival analysis and frailty models have continued to engross the field of statistics. Many varied applications of these methods have been shown in many different areas of science and even in outside fields like econometrics. The ability to utilize these models in computing has also taken off and has rapidly been expanding.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Usha Govindarajulu: Conceptualization; methodology; writing-original draft; writing-review and editing. **Ralph D'Agostino Sr.:** Conceptualization; methodology; writing-original draft; writing-review and editing.

ORCID

Usha S. Govindarajulu https://orcid.org/0000-0001-5864-7403

REFERENCES

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, *8*(8), 907–925.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.

Breslow, N. E., & Day, N. E. (1987). *Statistical methods in cancer research volume II: The design and analysis of cohort studies* (Vol. 82). Lyon: IARC Scientific Publication.

Chen, M.-H., & Ibrahim, J.,. D. S. (2002). Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis*, *80*(1), 101–126.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiologicalstudies of familial tendency in chronic disease incidence. *Biometrika*, *65*(1), 141–151.

Cook, R., & Lawless, J. (2002). Analysis of repeated events. *Statistical Methods in Medical Research*, *11*(2), 141–166.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B.*, *34*, 187–202.

Duchateau, L. J., & Janssen, P. (2008). *The frailty model*. New York: Springer.

Emura, T., Nakatochi, M., Murotani, K., & Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research*, *26*(6), 2649–2666.

Fleming, T.,. R., & Harrington, D. (1991). *Counting processes and survival analysis*. New York: John Wiley & Sons.

Georges, P., Arnaud-Gilhelm, L., Nicolas, E., Quibe, l. G., & Roncalli, T. (2001). *Multivariate survival modelling: A unified approach with copulas*. https://doi.org/10.2139/ssrn.1032559

Goethals, K., Jannsen, P., & Duchateau, L. (2008). Frailty models and copulas: Similarities and differences. *Journal of Applied Statistics*, *35*(9), 1071–1079.

Govindarajulu, U., Glickman, M., & D'Agostino, R. B., Sr. (2007). Modeling frailty as a function of observed covariates. *Journal of Statistical Theory and Practice*, *1*(1), 117–135.

Govindarajulu, U. S., Lin, H., Lunetta, K. L., & D'Agostino, R. B., Sr. (2011). Frailty models: Applications to biomedical and genetic studies. *Statistics in Medicine*, *30*(22), 2754–2764.

Ha, I. D., Jeong, J.-H., & Lee, Y. (2017). *Statistical modelling of survival data with random effects*. New York: Springer-Verlag.

Hogan, J. W., & Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, *16*(1–3), 239–257.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis; regression: Modeling of time-to-event data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Jones, G. (2004). Markov chain Monte Carlo methods for inference in frailty models with doubly-censored data. *Journal of Data Science*, *2*, 33–47.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley and Sons.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: John Wiley and Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation of incomplete observations. *Journal of American Statistical Associtation*, *53*(282), 457–481.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, *48*(3), 795–806.

Kleinbaum, D. G., & Klein, M. (2011). *Survival analysis: A self-learning text* (3rd ed.). New York: Springer Verlag.

Kosorok, M. R., Lee, B. L., & Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics*, *32*(4), 1448–1491.

Lin, D. Y., & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of American Statistical Association*, *84*, 1074–1078.

Liu, L., Wolfe, R. A., & Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, *60*(3), 747–756.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, *50*(3), 163–170.

Nduati, R., John, G., Mbori-Ngacha, D., Richardson, B., Overbaugh, S., Mwatha, A., ... Kreiss, J. (2000). Effect of breastfeeding and formula feeding on transmission of HIV-1: A randomized clinical trial. *Journal of the American Medical Association*, *283*(9), 1167–1174.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sorensen, T. I. (1992). A counting process approach to maximum likelihood estimation of frailty models. *Scandinavian Journal of Statistics*, *19*, 25–43.

Prentice, R. L., Williams, J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, *68*, 373–379.

Price, D. L., & Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, *20*(9–10), 1515–1527.

Rizopolus, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Boca Raton, FL: Chapman and Hall/CRC.

Rondeau, V., Schaffner, E., Corbiere, F., Gonzalez, J. R., & Mathoulin-Pelissier, S. (2013). Cure frailty models for survival data: Application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Statistical Methods in Medical Research*, *22*(3), 243–260.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.

Staniswalis, J. G., Thall, P. F., & Salch, J. (1997). Semiparametric regression analysis for recurrent event interval counts. *Biometrics*, *53*(4), 1334–1353.

Sy, J. P., & Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, *56*(1), 227–236.

Therneau, T., & Grambsch, P. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

Tsiatis, A. A., DeGruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, *90*, 27–37.

Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, *88*, 447–458.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454.

Wienke, A. (2010). *Frailty models in survival analysis*. CRC Press LLC.

How to cite this article: Govindarajulu US, D'Agostino Sr. RB. Review of current advances in survival analysis and frailty models. *WIREs Comput Stat*. 2020;e1504. https://doi.org/10.1002/wics.1504