METHODOLOGICAL REVIEW

Modeling Medical Prognosis: Survival Analysis Techniques

Lucila Ohno-Machado

Decision Systems Group, Brigham and Women's Hospital, Health Science and Technology Division, Harvard Medical School and Massachusetts Institute of Technology, 75 Francis Street, Boston, Massachusetts 02115

Received January 14, 2002; published online July 2, 2002

Medical prognosis has played an increasing role in health care. Reliable prognostic models that are based on survival analysis techniques have been recently applied to a variety of domains, with varying degrees of success. In this article, we review some methods commonly used to model time-oriented data, such as Kaplan-Meier curves, Cox proportional hazards, and logistic regression, and discuss their applications in medical prognosis. Nonlinear, nonparametric models such as neural networks have increasingly been used for building prognostic models. We review their use in several medical domains and discuss different implementation strategies. Advantages and disadvantages of these methods are outlined, as well as pointers to pertinent literature.

Key Words: survival analysis; neural networks; Cox proportional hazards; prognosis; machine learning.

1. INTRODUCTION TO MEDICAL PROGNOSIS

We define medical prognosis as an estimate of cure, complication, recurrence of disease, level of function, length of stay in health care facilities, or survival for a patient or group of patients. Prognostic information can (1) help medical researchers to identify certain patterns of disease progression; (2) assist health maintenance organizations and other members of the health care industry in predicting the needs of their served populations and, consequently, allocate resources accordingly; and (3) serve as an additional aid for patients to understand and make informed decisions about their conditions. Prognosis is therefore an important part of medical care, since it influences decisions regarding individual preferences, treatments, and resource allocation [1]. For example, patients who are likely to have an ominous prognosis may opt for aggressive new therapies; physician's decisions about transferring patients into and out of intensive care units (ICUs) may be determined, among other things, by each individual's probability of survival.

A common misconception is that medical prognosis relates exclusively to the prediction of patterns of disease progression in the absence of treatment. It is currently almost impossible to obtain such data, since in most cases patients will receive some form of therapy. Therefore, patterns of disease progression are often reported and predicted in relation to what is considered the “standard” therapy for that condition. Given the large variations in
“standard” therapy, data for prognostic studies are very difficult to collect, and the most reliable collections come from randomized clinical trials, usually performed in multiple academic medical centers.

The increasing availability of computers in private practices, clinics, and hospitals, the collection of structured data in electronic format, and the recent development of machine learning models that can be built to recognize certain patterns of disease progression make it possible (at least in theory) to construct specific prognostic systems that take into account not only general information about a patient’s diagnosis, but also patient-specific information, such as demographics, past medical history, current treatments, clinic-related information, such as personnel skills and available facilities, and, more recently, gene expression levels at the tissue level. Systems built with this specific information may have higher chances to be accurate for a particular patient. The future routine use of such systems has the potential to improve medical care substantially.

Among a myriad of techniques that can be used to model prognoses, there is a subset that uses information from data to “learn” models that should be generalizable to new cases. Those that fall under the umbrella term “survival analysis” are particularly important given their popularity in the medical domain. For example, several models have been built to classify patients into prognostic categories: the APACHE [2] scores for estimating the probability of death in ICUs, and the Framingham risk models [3] for estimating the risks of developing heart disease are based on logistic regression models. Formal testing of these prognostic tools and comparison with other models have been performed by several authors in specific populations. Other types of survival analysis techniques have been used for analogous purposes, and will be the object of this review.

It is important to note that other types of score models have been used to model medical prognosis, such as the Glasgow scores for coma, and the various staging systems for cancer: these are different from all the survival analysis methods covered in this article, in that they are constructed manually and then correlated with experimental prognostic data (unlike the survival analysis techniques, which are learned from the data directly). There is another class of techniques that are built like these scores (i.e., by manually modeling understanding of disease processes, physiology, or clinical knowledge, and then validating the performance of the model using experimental data) that are beyond the scope of the paper.

2. MODELING TIME UNTIL AN EVENT USING EXISTING DATA

Prognostic tasks can be divided into ones that (1) involve prediction for a single point in time and (2) involve time-related predictions. Included in the first category are prognostic tasks such as estimation of outcomes in a single time point within fixed time limits (e.g., survival in 5 years, extensively used in the oncology literature) or outcomes within variable (but usually well defined) time periods (e.g., survival in the intensive care unit of a hospital). Included in the second category are prognostic tasks that predict prognoses at several points in time, for example, the probability of coronary disease development during the life of an individual. Outcomes that are modeled in prognostic tasks can be either discrete (e.g., death) or continuous (e.g., survival, length of stay in the intensive care unit). As we will review in this article, the choice of model is dependent on both the specification of the prognostic task and the nature of the outcome. Before we discuss these items, however, it is necessary to introduce some basic concepts in modeling time-oriented data. For clarity, we distinguish two ways to model time until an event: modeling outcomes in a single interval of time (e.g., death within 1 year, development of disease within 10 years) or in multiple dependent intervals (e.g., yearly survival estimates covering a period of 10 years). In the former case, we can construct single-point models and, in the latter, we construct multiple-point models.

2.1. Modeling a Single Point versus Multiple Interdependent Points in Time

Single-point prognostic models are aimed at producing an estimate of the probability for the occurrence of the event of interest before a certain time (e.g., 50% probability of survival in 10 years) or an estimate of time elapsed until the event of interest (e.g., 120 days until death). A single estimate is produced by these models.

There are several problems with single-point estimation. A single-point estimate for a certain time limit may be misleading if its interpretation is extended for longer terms. Two different patients may have a 10-year survival estimate of 50%, but the disease in one of them may progress much slower than the other after that time (Fig. 1). If a single-point model is used, care should be taken in order not to improperly extrapolate its estimates to longer intervals.

For example, an isolated single-point estimate of survival may be useful for certain purposes, but provides no information on whether disease development seems to be fast or
The hazard function is defined as the probability that an individual will die at a certain time, conditioned on his survival up to that time, and denotes the instantaneous death rate [4]. It can be derived from the survival function and it can be represented as

$$h(t) = \lim_{\Delta t \to 0} \frac{P[death(t, t + \Delta t)]}{\Delta t}.$$  

where $death(x,y)$ is death in the interval starting at $x$ and ending in $y$ for individuals aged $x$. The cumulative hazard function is related to $S(t)$ by

$$H(t) = -\log S(t).$$

Survival analysis models, such as actuarial life tables [5], product-limit estimators [6], Cox proportional hazards [7], and fully parametric models [5], produce estimates for the survivor function and the hazard function. Parametric methods of survival analysis require specification of a probability density function for estimating these functions. All these methods have been used more frequently to explain the progression of the disease for known cases than to predict survival for new cases. This distinction is important because models that explain the progression of disease in a specific “training set” may not be generalizable to previously unseen “test sets.”

For predicting survival for new cases, logistic regression and neural networks have been more commonly applied.

2.2. Censoring

Survival analysis is confronted with several sources of difficulties. One of them is the possibility that some individuals may not be observed for the full time since the beginning of the disease up to the event being studied [7]. This type of missing data is called censored data and is exemplified in Fig. 2. Censored data may represent cases that were studied for the duration of the study, and whose survival was at least as long as the study, but whose final event time is unknown. This is called Type I censoring. Censored data may also represent the cases that were lost to follow-up after a certain time. For example, a patient may have been followed and is still alive after 5 years, but is then lost to a data collection system. His data will be useful to construct models that predict survival in 5 years, but not useful for models that predict survival beyond that. This is called Type II censoring [8].
conditioned on his or her survival at time \( t - 1 \). The survivor function for actuarial life tables and product-limit estimators is

\[
S(t) = \prod \left( \frac{n_j - d_j}{n_j} \right),
\]

where \( d_j \) is the number of deaths in interval \( j \), \( n_j \) is the number of individuals at risk, and the product is for all intervals from 1 to \( j \) \[4\]. In the actuarial method, \( n_j \) is the average number of individuals at risk; in the Kaplan–Meier method, \( n_j \) is the actual number of individuals at risk. Actuarial life tables and product-limit estimators differ in the way that time intervals \( j \) are built. The former model predefines intervals of equal duration and groups deaths in those intervals. The latter model builds one interval for each death, and therefore does not cause loss of information. An example of a Kaplan–Meier survival curve is shown in Fig. 3. Both models can handle censored data.

Life tables and product-limit estimators are good for describing survival of a group of individuals. Actuarial life tables and product-limit estimators produce survivor estimates for different strata and require arbitrary discretization of continuous variables (i.e., we can compare patients over 45 with those under 45 years of age, but this method cannot model the whole range of ages). Information regarding covariates not used for stratification is therefore not taken into account. Survival curves can therefore be built for a group of patients who share a given characteristic with respect to a particular variable. Comparison of two different survival curves produced for two strata is carried out using log-rank [9], the Mantel–Haenszel method [10], or Breslow tests [11]. Although comparisons with more than one variable are possible (e.g., age and gender), the number of patients in each resulting group must not fall below a certain minimum, otherwise a reliable survival curve cannot be created. These
models are good for data visualization and making comparisons based on one of two variables, but they are not appropriate for multivariate analysis.

4. REGRESSION MODELS

4.1. Cox Proportional Hazards

The Cox model is a multivariate semiparametric regression model that allows modeling of continuous covariates and involves the assumptions that there is a simplifying transformation of the initial data and that the hazards for the different groups are proportional. The model assumes that any baseline hazard and hazards for individuals with certain variable values are multiples of that baseline, as in

\[ h_i(t) = h_0(t)e^{x_i\beta}, \]

where \( h_i(t) \) is the hazard for individual \( i \) at time \( t \), \( h_0(t) \) is the baseline hazard at \( t \), \( x \) is the vector of explanatory variables, and \( \beta \) is the vector of coefficients for each variable. Figure 4 shows survival curves derived from a Cox proportional hazards model. Note that in the Cox model the survival curves may not cross. For problems in which the assumption of hazard proportionality does not hold (such as the one illustrated in Fig. 1), models of survival that allow nonproportional hazards are needed.

The main advantage of the Cox model over parametric models is that the baseline hazard does not need to be known in order to calculate the parameter \( \beta \) (and therefore to determine the importance of each covariate) assuming that the proportionality assumption holds true. Furthermore, it can approximate parametric models such as the Weibull and the exponential, so there is no need to assume a certain hazard function to calculate \( \beta \). This model is consequently very popular for describing how covariates influence survival in known cases because it will work for any baseline hazard, as long as the proportionality assumption holds (i.e., survival curves do not cross).

The Cox model is not commonly used to perform prognosis on new cases, but rather to characterize disease progression on existing cases, by revealing the importance of covariates. This model is therefore extremely popular. However, in order to make predictions of survival for new cases, taking into account the variables in \( x_i \) and the estimated parameter \( \beta \), a baseline hazard must be estimated. This estimation represents no trivial task and the choice of the wrong baseline hazard can change the results of the predictions dramatically. The baseline hazard can be derived from survivor functions, calculated empirically such as the ones in actuarial life-tables or product-limit estimators, or parametric models. If the latter are used, the model loses the property of being semiparametric. If a Weibull model is assumed, the baseline is characterized as

\[ h_0(t) = \lambda \gamma (\lambda t)^{\gamma - 1}, \]

where \( \gamma \) and \( \lambda \) are the shape and scale parameters, respectively.

Other parametric models, such as the exponential, can also be used for this purpose.

The Cox proportional hazards model is a multiple-point model, as estimates of survival for a particular patient are conditionally calculated relative to a baseline hazard. There are several examples of its use in different medical domains to model: survival with primary biliary cirrhosis (13), survival with CHD (14), development of CHD (15), survival with AIDS (16), survival with primary cancer of the lung (17), survival with breast cancer (18), and survival with ovarian cancer (19). Except for (16), these applications were related to the application of the model to known data, and not the prediction of survival for unknown data. The Cox proportional hazard model can handle censored data.

4.2. Logistic Regression Models

Logistic regression is the most used supervised algorithm for binary classification in medicine (10). A multiple logistic
regression model (in which the variables are represented as a vector $X$ of $n$ variables $x_1, x_2, \ldots, x_n$) can be written as

$$\text{logit}(p) = \log[p/(1 - p)] = \text{intercept} + \beta X.$$ 

The intercept is a constant, $\beta$ is the vector of coefficients for the variable vector $X$, and $p$ represents the estimated probability that a particular case belongs to the category assigned as “1” in the gold standard. The equation can be rearranged to

$$p = 1/[1 + \exp(- (\text{intercept} + \beta X))].$$

Simple logistic regression models produce estimates of probabilities of a certain dichotomous outcome, given specific values for the independent variables. These are single-point models and can therefore handle censored data. In survival analysis, the model has been used to estimate the probability of survival at each time point taking into account the variables $x$. As other single-point models, there is no guarantee that the probabilities produced for different points are consistent with each other.

In the pooled logistic regression model, data obtained from one individual may be used more than once as inputs to the logistic regression. For example, if data from a patient were collected at intervals $I_1$, $I_2$, and $I_3$, the researcher may use the three different records as inputs to the logistic regression models. This model is used in some of the Framingham risk models.

The APACHE score [20] for determination of disease severity based on prognostic survival is a good example of a logistic regression model that has been translated into routine care. It takes into account several patient parameters to determine the probability or survival in the ICU. There are several examples in the literature of logistic regression models used for prognosis. Some models deal with prognosis during hospitalization or a variable (but specified) time horizon and produce estimates of survival in ICU [20, 21], success of angioplasty [22–24], and functional outcomes for patients in rehabilitation facilities [25].

Examples of other models, in which the time horizon is open, include survival with primary cancer of the lung [17], mortality after myocardial infarction [26], outcomes of epilepsy surgery [27], early and late complications of stents [28], in-hospital pneumonia mortality [29], and in-hospital death due to alcoholic liver disease [30].

Other regression models have been constructed for continuous outcomes. For example, linear regression was used to model survival days, length of stay in hospital, costs in a rehabilitation facility [31], and volume of hemorrhagic blood loss under certain surgical conditions [32]. Quadratic discriminant analysis was used to predict success of weaning patients from ventilators [33]. Regression trees were used to model survival and presence of complications after coronary intervention procedures [34].

With the notable exception of the APACHE model, it is not common to encounter rigorous evaluations of the predictive performance of these models in previously unseen cases. A great advantage of using dichotomous logistic regression models is that it is easy to use them to predict the outcome in new cases and they are easily found in any off-the-shelf statistical package.

5. NEURAL NETWORKS

Neural networks constitute good alternatives for the predictive methods discussed above. Neural networks are computational models that may be used in the same tasks as linear or quadratic regression models, logistic regression, or Cox proportional hazards models.

Feed-forward neural networks can be seen as analogous to regression models, in which covariates are called “inputs,” coefficients are called “weights,” and the outcome variable is called “output” or “target.” In graphical representations such as the one in Fig. 5, the inputs and outputs are represented by circles, and the weights are represented by the connecting lines. Note that there may exist a layer of intermediate circles, which are called “hidden nodes.” This hidden layer allows neural networks to model nonlinear relationships between inputs and outputs, and can be thought to have a role similar to that of interaction terms in regression. The main difference is that the nature of these interaction terms is not explicit, and therefore not amenable to interpretation.

![FIG. 5. Single-point estimate neural network model that predicts the development of coronary heart disease (CHD) at time $t_0$.](image)
Neural networks are nonparametric models that are especially useful in situations where modeling of nonlinear relationships (or classification of nonlinearly separable data) are necessary. Although there seem to be actually few problems in medicine that require such a capability, there is no way to know in advance whether a classification problem contains data that are linearly separable in the prediction variable. Neural networks have been used for medical diagnosis in the past two decades, but their use for prognosis has been more limited. The backpropagation algorithm [35] has been used extensively to estimate parameters in neural networks, but several other algorithms exist. Some authors have compared neural networks to several statistical classification methods and outlined the similarities in scope and purpose of neural networks and logistic regression models [36]. In exchange for their flexibility in classification of nonlinearly separable data, neural networks present the disadvantage of easily overfitting the training data, a feature that translates into poor predictive ability with regard to previously unseen cases. It is actually precisely because certain architectures of neural networks are so flexible that they overfit (or memorize) the training data! Although some strategies exist to avoid overfitting, some involve the use of part of the training set for “tuning” the networks, leaving fewer cases for the parameter estimation. Therefore, neural networks can produce results that are better, the same, or worse than other models such as logistic regression. Furthermore, because these models are complex, their interpretation is severely impaired, and they are often considered to be “black boxes.” Because it is not possible to determine which model (e.g., logistic regression, neural network) will be best for a given data set, it is recommended that all of them be tried in case interpretability of the resulting model is not a priority.

A few applications of neural networks that model survival as a continuous outcome exist: prediction of length of stay in hospital in the ICU setting [37, 38], length of stay for treatment of rib fracture [39], length of stay in a rehabilitation facility [31], length of stay in acute pancreatitis [40], and hemorrhagic blood loss in experimental surgery [32]. Most of these applications did not compare results with analogous and simpler linear or quadratic regression models. This type of model is not popular because it is less related to survival analysis models such as the ones presented before, in which the dichotomous concept of failure (or death) is fully incorporated. We describe next neural network models that have been more widely applied.

5.1. Single-Point Models

There are several ways to build neural networks for survival analysis. A simple way is to build single-point models, in which predictions within a specified period are modeled, as explained in the previous section (e.g., survival within 1 year of surgery). Single-point neural network models are simple and efficient in situations where we are only interested in prognoses at a specified time limit. In cancer studies, it is customary to report survival rates within 5 years of treatment. In public health studies involving the development of certain diseases or conditions, it may be interesting to look at event occurrence within 10 years. If that is the case, a simple network can be constructed, as illustrated in Fig. 5.

An aggregate neural network that produces multiple-point estimates of survival, such as the one shown in Fig. 6, can be used to produce survival curves (analogous to aggregating several logistic regression models). This network, however, is not able to use predictions for a given interval to enhance predictions for another interval, unless explicit weights are built among the outputs.

Utilizing aggregates of single-point estimates to predict individual survival and to delineate patterns of disease progression may yield spurious results, including nonmonotonic survival estimates, as illustrated in Fig. 7.

Spurious curves are created, among other reasons, because predictions tend to be less accurate in certain intervals [41], and no provisions for accounting for dependency among the single-point estimates are made in the isolated models.

Some examples of single-point models that are found in the medical literature have addressed: valve-related complications in heart disease [42], adequacy of admission to the psychiatry ward [43], outcomes of liver transplantation [44–
allows the construction of survival curves if necessary. It has
the advantage of being very simple, since a single network
models makes predictions for all intervals. It involves con-
siderable preprocessing of data, with selective duplication
cases that have longer survival. For example, if patient
A lives 5 years and patient B lives 1 year, the input data set
may contain five copies of patient A (one for year 1, one
for year 2, etc.), but only one copy of patient B. Ravdin has
developed a method to account for this bias in the final
prediction. This architecture uses a single model to map
inputs to outputs (patient features to predictions of survival).
The quality of predictions for different time points varies
according to the number of cases with a long enough follow-
up and the distribution of cases for a specific time point.
This network makes a single universal model to make predic-
tions for all intervals and shares some of the problems out-
lined above for models based on aggregates of single-
point estimates.

5.2. Multiple-Point Models

As we discussed previously, if predictions for several
intervals are sought, a simple aggregation of networks can
produce unrealistic estimates. It is therefore desirable to
build networks that can account for dependencies in several

FIG. 8. Coding time as an input in a model that makes prognosis
corony heart disease (CHD). In the training set, patient information
at each known time interval $t_i$ is entered. The input node "$i$" represents
the time elapsed until that interval. For new patients, the desired time
for prediction is entered, and the network produces the prognosis for
that period. This architecture can handle censored data.

neural networks, dependencies among time point predictions are explicit-
ly represented, resulting in monotonic survival curves. In this example,
predictions for year 3 are entered in the model that predicts survival
in year 4. This architecture can handle censored data.
The recognition of patterns of disease progression can be made by neural networks that take into account dependencies among output variables. A combination of neural networks [41] was used to obtain monotonic survival curves in prognostic tasks such as survival with AIDS [16], and survival without coronary heart disease [63]. In this system, shown schematically in Fig. 9, predictions from time \( t_a \) can be improved if predictions on time \( t_b \) are provided as inputs, and so on. A temporal pattern can be delineated this way. Sequential neural network systems are constructed incrementally. In each step of the sequence, predictions for one time point are produced by a neural network. These predictions are passed forward to other networks in the system. One or more networks may provide predictions that become inputs for other networks in the sequence. The result is a chain or a hierarchy of neural networks, as illustrated in Fig. 9.

6. WHICH MODEL IS BEST?

We summarize in Table 1 the main features of using single-point versus multiple-point models, and where each regression or neural network model fits. Extensive empirical comparison of all models in the same data set has not been reported, although there has been increased interest in using neural networks for prognostic tasks using real data sets. A few of these reports contain some form of comparative evaluation involving neural networks and more traditional models. The fields of application have been varied, and results highly dependent on the quality of training data and evaluation methods used. The variety of results indicate that indeed there are no theoretical justifications to choose one method over the other, especially when interpretability is not at stake.

Neural networks have been shown to perform better than other models in predicting: survival after surgery for non-small cell lung carcinoma [56], development of coronary heart disease [15], hemorrhagic blood loss [32], survival, cost, length of stay in a rehabilitation facility [31], death in ICU [21], breast cancer survival (vs a histology-based system) [67], and colon cancer survival [68]. In other experiments, neural networks showed no advantage over other models in predicting: breast cancer survival [69], ovarian cancer survival [19], breast cancer recurrence [66], in-hospital pneumonia mortality [29], coronary artery bypass graft surgery success [22], and in-hospital death due to alcoholic liver disease [30].

In analyzing the results from the studies mentioned above,
it is important to keep in mind that these applications use previously collected data to build predictive models for new cases and are limited, among other things, by the quality of the original data and the intrinsic characteristics of the methods employed. Publication bias is also an important factor, so negative results may have been underrepresented.

5. SUMMARY

Although the main focus of pioneer applications of computer-based medical decision support was the diagnosis and choice of treatment for diseases, more recent applications have been built to assess the prognosis for a certain patient profile. These multivariate survival analysis applications utilize modeling techniques such as linear and logistic regression models, modified versions of the Cox proportional hazards method, and neural networks. The advantages and disadvantages of using different methods for predicting survival have seldom been tested in large collections of data sets, and it is unlikely that one method can always outperform the others, especially when predictive performance is not the only index for comparison.

Understandably, it is often the case that other features of a model, such as its interpretability, easy implementation in clinical settings, and acceptability by the medical community intervene in favor of a method that may not have the highest predictive performance. Small differences in predictive performance may indeed not be as important as the ability to interpret or use a model in a clinical setting. Although the use of artificial data sets facilitates the control of data for some of these comparative studies, the use of real data sets provides more usable and convincing information.

In real data sets, censoring of observations, other types of missing data, and noise are frequently present and difficult to control. A good survival predictor must be able to deal with these obstacles. In this article, we have reviewed some existing methods for modeling prognosis using survival analysis techniques. A detailed presentation of most of the concepts and models described here can be found in [8].

ACKNOWLEDGMENTS

This work was funded in part by Grant R29LM06538-01 from the National Library of Medicine. The author thanks two anonymous reviewers for their insightful comments on a previous version of the manuscript.

REFERENCES


