Multistate survival models as transient electrical networks

Ronald W. Butler
Department of Statistical Science
Southern Methodist University
Dallas, TX 75275 U.S.A.
rbutler@smu.edu

Douglas A. Bronson
Neptune and Company
Lakewood, CO 80215 U.S.A.
dbronson@neptuneinc.org

November 2, 2009

Summary

In multistate survival analysis, the sojourn of a patient through the various clinical states is shown to correspond to the diffusion of 1 coulomb of electrical charge though an electrical network. The essential comparison has differentials of probability for the patient correspond to differentials of charge and equates clinical states to electrical nodes. Indeed, if the death state of the patient corresponds to the sink node of the circuit, then the transient current that would be seen on an oscilloscope as the sink output is equivalent to the probability density for the survival time of the patient.

This electrical circuit analogy is further explored by considering the simplest possible survival model with two clinical states - alive and dead. The corresponding states of a circuit are its source and sink nodes. For the survival model, if the patient's lifetime is subject to independent right censoring and left truncation, then Kaplan-Meier is the appropriate estimate for survival time free from censoring risk and truncation. When appropriate analogs to censoring and truncation are incorporated into an electrical circuit, then the sink output that would be seen on an oscilloscope is also the Kaplan-Meier mass function.

An important consequence of this electrical analogy is the support it provides for the use of maximum likelihood as a method for statistical inference. The dynamics of current flow gives a fundamentally different motivation for the Kaplan-Meier estimator which would normally be motivated as a nonparametric maximum likelihood estimator.

A competing risks setting has multiple death states and corresponds to a circuit with multiple sinks. Again, after adjusting for censoring and truncation, Kaplan-Meier mass functions are the outputs at the sink nodes for the corresponding circuits.

If covariates are present, then the electrical analogy provides for an intuitive understanding of partial likelihood and the various baseline hazard estimates that are often used with the proportional hazards model.

Some key words: Competing risks; Electrical network; Kaplan-Meier; Left truncation; Multistate survival; Proportional hazards; Right censoring; Self-consistency; Semi-Markov.

1 Introduction and overview

Multistate survival models are used to describe the sojourn of a patient through various clinical states of an illness. These states typically consist of transient states and absorbing (death) states so that the sojourn is a first passage to one of the absorbing states. Rather than thinking of the patient occupying a single state at any time, it is perhaps more beneficial to consider the diffusion of probability for the patient through the clinical states so that, the patient is considered occupying all states (with certain probabilities) all of the time. The patient becomes a "virtual" patient defined in terms of occupancy (interval transition) probabilities. If probability is equated to charge, then the multistate survival process may be considered as a diffusion of 1 coulomb of electrical charge through an electrical network. The essential comparison has differentials of probability for the patient correspond to differentials of charge and equates clinical states to electrical nodes. Indeed, under this analogy, if the death states of the survival model correspond to the sink nodes in the circuit, then the transient current that would be seen on an oscilloscope at any particular sink output is equivalent to the density of the subdistribution of the patient dying from that particular cause. Such subdistribution densities integrate to give the absorption probabilities for their associated death causes or, in the electrical analog, the total charge that is absorbed at that sink node.

Section 1 reviews the basics of semi-Markov processes that have traditionally been used to represent multistate survival models. The emphasis is on developing the

relevant tools related to determining sojourn times of patients and interval transition probabilities. Section 2 builds upon the semi-Markov structure by developing the electrical network analogy just described and finding electrical analogs to sojourn time distributions as impulse response functions.

Previous analogies between stochastic processes and electrical networks were introduced by Kelly (1978) when considering the dynamics of reversible Markov processes. Kelly introduced two analogous electrical models. Our model agrees with his first and perhaps least considered model in which probability is equated with electrical charge. In his first model, Kelly (1978) shows that for certain Markov processes, the forward equations given in his equation (1.20), which describe the transient behavior of the process, are equivalent to Kirchhoff's equations for current flowing in an analog circuit with resistance between nodes (states) and capacitors grounding out the nodes. The particular processes for which this holds are processes which, when started in their stationary distribution, satisfy local balance equations so that they are reversible.

In both of Kelly's electrical analogies, local balance is critical in establishing the conductivities and hence resistances between nodes (states) of the analogous electrical network. Kelly's second electrical analogy equates probability to electrical potential (voltage) and also assumes processes are reversible in their equilibrium condition. Doyle and Snell (1984) build upon this potential analogy which, over the years, has resulted in this particular model becoming the better known probabilistic analogy. This second approach, however, is not considered below.

Of course survival models are for the most part transient processes and the analogy developed below allows for transient processes that are not constrained by the assumptions of stationarity and reversibility used by Kelly (1978) and Doyle and Snell (1984). The Markov assumption is also relaxed to allow consideration of semi-Markov models traditionally used in multistate survival analysis.

Our circuitry analogy for transient semi-Markov processes is further explored by considering the simplest possible survival model that has two clinical states - alive and dead. The corresponding states of an electrical network are its source and sink nodes. If patients' lifetimes are subject to independent right censoring and left truncation, then Kaplan-Meier would be an appropriate estimator of survival that is free from censoring risk and truncation. By indulging upon this electrical network analogy, appropriate censoring states and truncation options can be built into an "empirical" electrical network based on the data. When this is done, the impulse response function that would be would be seen on an oscilloscope attached at sink output, is the Kaplan-Meier mass function. The essential idea, is that the superpositioning of current flows in the empirical electrical network sets up a transform domain version of the self-consistency equations introduced by Efron (1967). Inversion of these transforms leads to Efron's self-consistency equations in the time domain whose solution is both the Kaplan-Meier mass function as well as the impulse response function at the death node.

When considering proportional hazards models, the electrical analogy provides simple yet intuitive interpretations for partial likelihood and also for the various baseline hazard estimators. When covariate value u is considered in this setting, survival time X_θ with $\theta = \exp(\beta^T u)$ has survival function $S_0(t)^\theta$ where $S_0(t)$ is baseline survival. Within the electrical framework, the patient with lifetime X_θ is assigned charge θ rather than 1. Partial likelihood is based upon hazard probabilities that are recorded in terms of the fraction of total charge at risk at each of the death times.

The competing risks setting is a simple extension of the single risk setting to which the electrical analogy may be extended by having multiple sink nodes. All the ideas discussed carry through into this setting. For example, the impulse response

functions at the various sink nodes turn out to be Kaplan-Meier weights that are associated with the corresponding Kaplan-Meier subdistribution estimates. These output responses are a consequence of applying the superpositioning of current flows in an empirical electrical network to construct transform domain versions of some self-consistency equations whose solutions are Kaplan-Meier subdistributional mass functions.

A fundamentally important consequence of this electrical analogy is the support it provides for the use of maximum likelihood as a method of statistical inference. The Kaplan-Meier estimator is a nonparametric maximum likelihood estimator which adjusts for right censoring and left truncation. However, by equating patient transitions with the flow of current through an empirical electrical circuit designed to account for censoring and truncation, the Kaplan-Meier mass function is also obtained through an alternate motivation. This motivation uses superpositioning of current flows to justify self-consistency relationships that characterize the Kaplan-Meier estimator. Thus, the electrical analogy provides a separate and fundamentally different motivation for an estimator that would otherwise be motivated in terms of mathematical likelihood.

2 Semi-Markov survival models

Multistate survival analysis has traditionally used semi-Markov models to describe the movement of a patient afflicted with an illness through its various clinical states $S = \{1, \ldots, m\}$. In the simplest setting, a patient enters state 1 at baseline time 0 and moves from state to state while holding in each state for a random amount of time. Upon entering state 1, a semi-Markov model presumes that exit from state 1 is a competing risk situation for which the m-dimensional distribution $\mathbf{H}_1$ determines

the holding time and next state. If random vector $(H_{11}, \ldots, H_{1m})$ has distribution $\mathbf{H}_1$, then $j_1 = \arg\min_j\{H_{1j} : j \in S\}$ is the next state and the holding time in state 1, with destination j_1 assured, has cumulative distribution function (CDF)

$$P(H_{1j_1} \leq t \,|\, H_{1j_1} = \min_j H_{1j}\}.$$

Upon entering state j_1 at time H_{1j_1}, departure from state j_1 becomes another competing risk with distribution $\mathbf{H}_{j_1}$ that depends upon the holding state j_1 but which is otherwise independent of the past. The sojourn continues in this way through the various states of a semi-Markov model so that the sojourn itself can be thought of as a sequence of independent competing risk exits from the visited states.

From this description of a sojourn, one can see that the collection of m-dimensional exit distributions $\mathbf{H}_1, \ldots, \mathbf{H}_m$ from states determines the process dynamics. However, not all aspects of these distributions are estimable from sojourn data through S. Those aspects of $\mathbf{H}_i$ that are estimable can be characterized as the collection of subdistributions associated with exit from state i; see Miller (1981, §8.2). Subdistribution $C_{ij}(t)$ is defined as the probability of making the transition from state i to state j before duration t; an unbiased estimate from sojourn data would be the sample proportion of exit data from state i that pass to state j before duration t.

Subdistributions for exiting state i are more formally defined as

$$C_{ij}(t) = P\{H_{ij} = \min_k(H_{ik} : k \in S) \leq t\} = p_{ij}D_{ij}(t),$$

where

$$p_{ij} = P\{j = \arg\min_k H_{ik}\}$$

$$D_{ij}(t) = P\{H_{ij} \leq t \,|\, j = \arg\min_k H_{ik}\}.$$

Here, $\{p_{ij} : j \in S\}$ are exit probabilities from state i and $D_{ij}(t)$ is the holding time CDF in state i given passage to state j is assured.

While Miller (1981) focusses on the estimability of subdistributions, a more conventional and equivalent approach to competing risks, such as in Cox and Oakes (1984, §9.2), emphasizes estimability for the collection of cause specific hazards associated with exit from state i, or

$$h_{ij}(t) = P\{\text{passage } i \to j \text{ in } (t, t+dt) \,|\, \text{holding in } i \text{ for duration } t\}/dt$$

for $j \in S$. The equivalence of these approaches rests on the one-to-one correspondence between the set of exit subdistributions $\{C_{ij}(t) : j \in S\}$ and the set of cause specific hazards $\{h_{ij}(t) : j \in S\}$. The formal relationship, which may be inferred from Cox and Oakes (1984, eqn. 9.3) is

$$\begin{aligned} dC_{ij}(t)/dt &= P\{i \to j \text{ in } (t, t+dt) \cap \text{holding in } i \text{ for duration } t\}/dt \\ &= h_{ij}(t) \exp\left\{ - \int_0^t \sum_{k=1}^{m} h_{ik}(u) du \right\}. \end{aligned}$$

Subdistributions provide the link for making direct comparisons with electrical networks and will be the focus of this discussion.

2.1 Convolutions and Transmittances

Any distributional consideration of two successive state transitions entails convolving sets of subdistributions in the time domain. The even more complicated analysis of a sojourn through S thus becomes intractable unless these convolutions are considered in terms of the Laplace-Stieltjes transforms of the subdistributions or

$$T_{ij}(s) = \int_0^\infty e^{st} dC_{ij}(t) = p_{ij} \int_0^\infty e^{st} dD_{ij}(t) = p_{ij} M_{ij}(s).$$

In electrical engineering, function $T_{ij}(s)$ is a *transmittance* and is the product of a transition probability and a moment generating function (MGF). The $m \times m$ matrix function

$$\mathbf{T}(s) = \{T_{ij}(s)\} = \{p_{ij}\} \odot \{M_{ij}(s)\} := \mathbf{P} \odot \mathbf{M}(s)$$

is the transmittance matrix of the semi-Markov process and is the component-wise product of transition probability matrix $\mathbf{P}$ and $\mathbf{M}(s)$, a matrix of one-step MGFs. The transmittance matrix $\mathbf{T}(s)$ characterizes the semi-Markov process and also provides the means for analyzing and computing sojourn times over state space S.

2.2 Sojourn transmittances

The sojourn time X of a patient entering state 1 at time 0 and passing to an absorbing state m can be specified in terms of transmittance matrix $\mathbf{T}(s)$. Define

$$f_{1m}\mathcal{F}_{1m}(s) = E(e^{sX}1_{\{X<\infty\}}) \tag{1}$$

as the first-passage transmittance from state $1 \to m$, where $f_{1m} = P(X < \infty)$ and $\mathcal{F}_{1m}$ is the conditional MGF given $\{X < \infty\}$. A simple general relationship between (1) and $\mathbf{T}(s)$ has been given in Butler (2000) and is replicated in (5) below.

Before considering this general relationship however, consider the simple illness-death model shown in the flowgraph in Figure 1.

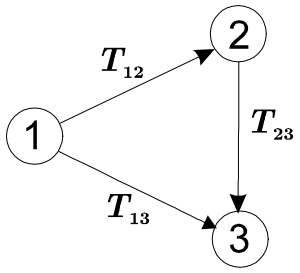

Figure 1. Illness-death model showing transmittances. States 1,2, and 3 represent good health, illness, and death respectively.

A patient starts in state 1 at time 0 with good health and eventually dies at time $(t, t + dt)$ in one of two distinct ways. Either the patient dies directly $(1 \to 3)$ with probability (w.p.) $dC_{13}(t)$ or becomes ill $(1 \to 2)$ in time $u \in (0, t)$ w.p. $dC_{12}(u)$ and subsequently dies in time $t - u$ w.p. $dC_{23}(t - u)$. Summing over the two distinct

paths and over all possible transition times $u \in (0,t)$ from $1 \to 2$ gives transmittance density

$$P\{X \in (t, t+dt)\} = dC_{13}(t) + \int_0^t dC_{12}(u) dC_{23}(t-u). \tag{2}$$

Since this is a proper density with $f_{13} = 1$, the first-passage transmittance is computed as the transform of (2) which gives

$$\mathcal{F}_{13}(s) = T_{13}(s) + T_{12}(s)T_{23}(s). \tag{3}$$

The example is quite simple because all states are progressive, meaning once a state has been entered, it cannot be reentered. The computation of $\mathcal{F}_{1m}(s)$ becomes more complicated with non-progressive states that can be reentered an arbitrary number of times. Non-progressive states result when feedback loops are present in the flowgraph. For example, if transmittance T_{21} is added to the illness-death flowgraph in Figure 1, the path $1 \to 2 \to 1$ constitutes a feedback loop whose presence allows for an arbitrary number of reentries into states 1 and 2. This leads to a countably infinite number of distinct paths from $1 \to 3$. Summing over all such paths and using the same argument used to derive (3) leads to the more complicated first-passage transmittance

$$\begin{aligned}\mathcal{F}_{13}(s) &= \sum_{k=0}^{\infty} \{T_{12}(s)T_{21}(s)\}^k T_{13}(s) + \sum_{k=0}^{\infty} \{T_{12}(s)T_{21}(s)\}^k T_{12}(s)T_{23}(s) \\ &= \frac{T_{13}(s) + T_{12}(s)T_{23}(s)}{1 - T_{12}(s)T_{21}(s)}.\end{aligned} \tag{4}$$

With additional states and more complex feedback patterns amongst the states, summing over all distinct paths from $1 \to m$ becomes intractable without general expressions for doing so. Butler (2000) provides such an expression in which (1) is determined from $\mathbf{T}(s)$ as

$$f_{1m}\mathcal{F}_{1m}(s) = \frac{(m,1) \text{ cofactor of } \mathbf{I}_m - \mathbf{T}(s)}{(m,m) \text{ cofactor of } \mathbf{I}_m - \mathbf{T}(s)} := \frac{(-1)^{m+1}|\mathbf{\Psi}_{m1}(s)|}{|\mathbf{\Psi}_{mm}(s)|} \tag{5}$$

where $\mathbf{\Psi}_{m1}(s)$ is the $(m,1)$ minor of $\mathbf{I}_m - \mathbf{T}(s)$, etc. Expression (5) is well-defined as the first-passage transmittance if S contains all relevant states to the sojourn $1 \to m$

and does not contain any irrelevant states, i.e. states that cannot possibly be transient intermediate states during the sojourn.

Expression (3) is easily derived using (5) with

$$\mathbf{T} = \begin{pmatrix} 0 & T_{12} & T_{13} \\ 0^* & 0 & T_{23} \\ 0 & 0 & 0 \end{pmatrix}. \tag{6}$$

Expression (4) results when 0^* is replaced by T_{21} in (6) to form a feedback loop.

In control theory and electrical engineering, first-passage transmittance $f_{1m}\mathcal{F}_{1m}(s)$ would traditionally be computed by using Mason's Gain Rule; see Phillips and Harbor (1996). If the two cofactors in ratio (5) are expanded in terms of their permutation sums and the non-zero terms in these sums are appropriately organized, then (5) can be shown to agree with Mason's Rule. A general proof of this is given in Butler (2001).

2.3 Interval transition probabilities

Let $Y(t)$ be the state of a patient at time t who enters state 1 at time 0. The interval transition probability functions are

$$\mathcal{P}_i(t) = P\{Y(t) = i\}$$

for $i \in S$. In the time domain, these functions are quite complicated convolutions of subdistributions, but their Laplace transforms (using kernel e^{-st}) are simple functions of $\mathbf{T}(-s)$. Pyke (1961, eqn. 4.3) has shown that

$$\mathcal{P}_i^*(s) = \int_0^\infty e^{-st}\mathcal{P}_i(t)dt = \frac{(i,1) \text{ cofactor of } \mathbf{I}_m - \mathbf{T}(-s)}{|\mathbf{I}_m - \mathbf{T}(-s)|} \times \frac{1 - \sum_{k\in S} T_{ik}(-s)}{s}. \tag{7}$$

Medhi (1994, §7.3.1) also provides a derivation.

As a simple example, consider computation of $\mathcal{P}_1(t)$ in the illness-death model. The value of $\mathbf{T}$ in (6) gives

$$\begin{aligned}\mathcal{P}_1^*(s) &= \frac{(1,1) \text{ cofactor of } \mathbf{I}_3 - \mathbf{T}(-s)}{|\mathbf{I}_3 - \mathbf{T}(-s)|} \times \frac{1 - \sum_{k \in S} T_{1k}(-s)}{s} \\ &= 1 \times \frac{1 - T_{12}(-s) - T_{13}(-s)}{s}, \end{aligned} \tag{8}$$

which is the Laplace transform of

$$1 - p_{12}F_{12}(t) - p_{13}F_{13}(t) = P(\text{Occupy state 1 at time } t).$$

The computation of $\mathcal{P}_2(t)$ in the illness-death model is less trivial, but it also provides a more general understanding of the factors in (7). This expression is

$$\begin{aligned}\mathcal{P}_2^*(s) &= (-1)^{2+1} \left| \begin{pmatrix} -T_{12}(-s) & -T_{13}(-s) \\ 0 & 1 \end{pmatrix} \right| \times \frac{1 - T_{23}(-s)}{s} \\ &= T_{12}(-s) \times \frac{1 - T_{23}(-s)}{s}, \end{aligned} \tag{9}$$

which is the Laplace transform for convolution

$$\int_0^t p_{12}F_{12}'(u) \times \{1 - F_{23}(t-u)\}du := \int_0^t E_2(u) \times H_2(t-u)du,$$

where

$$\begin{aligned} E_2(u)du &= P\{\text{Enter into state 2 at time } (u, u+du)\} \\ H_2(t-u) &= P\{\text{Hold in state 2 for } t-u \,|\, \text{Enter into state 2 at time } u\}. \end{aligned}$$

The same interpretation for (7) applies in the general setting of an m-state network with transmittance $\mathbf{T}$ and initial state 1 entered at time 0. The entrance rate $E_i(t)$ into state i at time t and holding probability $H_i(t)$ in state i have respective Laplace transforms

$$\frac{(i,1) \text{ cofactor of } \mathbf{I}_m - \mathbf{T}(-s)}{|\mathbf{I}_m - \mathbf{T}(-s)|} \qquad \text{and} \qquad \frac{1 - \sum_{k \in S} T_{ik}(-s)}{s} \tag{10}$$

that are the two factors of (7).

3 Electrical network analogy

Now consider the clinical states S of the survival model as the nodes for the analogous electrical network. A patient that arrives in state 1 at time zero becomes a Dirac function $\delta(t)$ input at node 1 that records current with units dependent on the scale of time for t. The total charge arriving in the network at instant $t = 0$ is therefore $\int_0^{0^+} \delta(t)dt = 1$ coulomb. The analogy equates probability 1, the certain arrival of the patient at time 0, with 1 coulomb of charge. From time 0 onward, the diffusion of probability for the patient through the clinical states is exactly the diffusion of charge through nodes of the electrical network. The fundamental idea is to dismantle probability 1 into differentials of probability which equate with differentials of charge flowing through nodes of the network. Whether considered as probability or charge differentials, this collection of differentials represents the "virtual patient".

3.1 Nodal charges over time

With probability as charge, then $\mathcal{P}_i(t)$ is both the probability the patient occupies state i at time t and also the total charge residing at node i at time t. The virtual patient occupies all states in S all of the time ($t > 0$) with positive probabilities, i.e. $\mathcal{P}_i(t) > 0$ for all $i \in S$, if $C'_{ij}(t) > 0$ for all $t > 0$ and all $i, j \in S$. Were the computation of $\mathcal{P}_i(t)$ to be performed in the time domain, it would require two layers of summation. The outer sum would be over all distinct finite-step paths from $1 \to i$. For each such path, the inner sum would convolve the path's associated subdistribution differentials over all allocations of the allotted time $(0, t]$ to states along the particular path for which $Y(0) = 1$ and $Y(t) = i$. Of course the flow of charge actually follows every possible finite-step path according to every possible allocation of finite time. However, working in the transform domain and computing $\mathcal{P}_i^*(s)$ instead eliminates the inner

layer of convolving since transmittances for finite-step paths get multiplied under convolution of their associated subdistributions. Indeed, were the right side of (7) to be fully expanded in terms of permutation sums for the determinants as well as for the division of determinants, formula (7) would lead to summation over all distinct finite-step paths from $1 \rightarrow i$.

3.2 Impulse response function

In first-passage transmittance (5), the probability f_{1m} that the patient reaches state m in finite time is the total charge that eventually reaches sink node m. If $f_{1m} < 1$, then X has a defective distribution that puts probability $1 - f_{1m}$ at ∞ which corresponds to charge that never reaches sink m.

In the survival model, MGF $\mathcal{F}_{1m}(s)$ is the transform for the density/mass function of sojourn time X given $X < \infty$ or $f_X(t)$. Product $f_{1m}f_X(t)$ is the arrival rate of sojourn probability at time t. For the electrical network, $f_{1m}f_X(t)$ is the arrival rate of charge or current arriving at node m at time t. A plot of current $f_{1m}f_X(t)$ versus t is the impulse response function for the Dirac input at node 1 which would be seen on an oscilloscope attached to node m. The transform of this impulse response function (with kernel e^{st}) is first-passage transmittance $f_{1m}\mathcal{F}_{1m}(s)$ which is often referred to as the transfer function at node m.

The computation of $f_{1m}\mathcal{F}_{1m}(s)$ in (5) assumes a flow of charge that follows every possible finite-step path from $1 \rightarrow m$ according to all possible allocations of finite time. If the right side of (5) were to be fully expanded using permutation sums for the determinants as well as for the division of determinants involved, then the formula would lead to a summation over all distinct finite-step paths from $1 \rightarrow m$ of the associated path transmittances, defined as the product of one-step transmittances along the path. As an example, see (3) for the simple illness-death model. Path

transmittances serve in place of a second layer of convolving which would be required for direct computation of $f_{1m}f_X(t)$ in the time domain as seen, for example, in (2). Transform analysis is preferred here because of its tractability and also because $\mathcal{F}_{1m}(s)$ is easily and accurately inverted by using saddlepoint methods to determine highly accurate approximations for $f_X(t)$ and its CDF; see Butler (2000 and 2007, ch. 13).

3.3 Limiting behavior

Limiting behaviors as $t \to \infty$ can be compared within the two interpretations of our model. If the semi-Markov model has transient and absorbing states only, then

$$\lim_{t\to\infty} \mathcal{P}_i(t) = \begin{cases} 0 & \text{if state } i \text{ is transient} \\ \alpha_i & \text{if state } i \text{ is absorbing,} \end{cases} \tag{11}$$

where $\alpha_i > 0$ is the absorption probability into state i when starting in state 1. For the electrical network, α_i is the total charge that accumulates in sink node i. In this transient setting, the limiting current into state i is $\lim_{t\to\infty} \mathcal{P}_i'(t) = 0$ for all $i \in S$. This may be shown by noting that

$$\lim_{t\to\infty} \mathcal{P}_i'(t) = \lim_{s\to 0}\, s \int_0^\infty e^{-st}\mathcal{P}_i'(t)dt = \lim_{s\to 0}\, s\left\{s\mathcal{P}_i^*(s) - \mathcal{P}_i^*(0)\right\} = 0. \tag{12}$$

Alternatively, in the case where S is finite and the semi-Markov process has stationary distribution $\{\pi_i : i \in S\}$, the limit in (11) is π_i the steady-state charge at node i; see, for example, Medhi (1994, §7.4). In this case, the limit in (12) is likewise 0 for all i since the net flow of current into state i in its equilibrium state is 0.

3.4 Summary

The underlying reasons why multistate survival models and electrical networks share a common mathematical analysis can be traced to double summation representations

for sojourn probability densities and impulse response functions as well as for interval transition probabilities and nodal charge distributions. A sojourn density (or impulse response function) from $1 \rightarrow m$ at time $(t, t + dt)$ is an outer summation, over all finite-step paths from $1 \rightarrow m$, of an inner summation that convolves randomly timed path steps so that first arrival in state m is assured at time $(t, t + dt)$. Thus first passage through a finite-state feedback network is essentially a countably infinite parallel connection of finite-step paths that are, in turn, series connections of randomly delayed steps. The two models, for a patient sojourn and the flow of electrical charge, treat parallel connections and series connections the same way (but for different reasons) and is why they share a common mathematical basis. The same explanation holds for the mathematical equivalence between interval transition probabilities of a patient and nodal charge distributions.

First consider the inner series connection associated with a particular finite-step path $1 \rightarrow 4 \rightarrow 2 \rightarrow m$. The path probability $p_{14}p_{42}p_{2m}$ and the MGF for path passage time $M_{14}(s)M_{42}(s)M_{2m}(s)$ are factors in the path transmittance regardless of whether the flow is probability or charge. For the patient, the product of one-step transmittances along the path, $p_{14}M_{14}(s) \times p_{42}M_{42}(s) \times p_{2m}M_{2m}(s)$ computes the probability of taking the path times the MGF for the time to traverse the path assuming independent holding times in the path states. For the flow of charge, the path transmittance is the product of two terms: the proportion of current following the path out of node 1 and the transform for a convolution that reflects the superpositioning of continuous current flow for all $t > 0$ through the series connection of the nodes in path $1 \rightarrow 4 \rightarrow 2 \rightarrow m$. The main point here is that the path transmittance is the correct transform in both instances: whether the patient moves from $1 \rightarrow 4 \rightarrow 2 \rightarrow m$ in only one particular time increment $(t, t + dt)$ or flows as current through the series connection during all time increments, as with the virtual patient.

Secondly, the outer summation of path transmittances over all distinct finite-step paths for $1 \rightarrow m$ is the proper computation whether considering patient movement or charge flow. To make this more clear, consider 3 parallel paths as in Figure 2 with transmittances $T_i(s) = p_i M_i(s)$ as Laplace-Stieltjes transforms of subdistributions $p_i F_i(t)$.

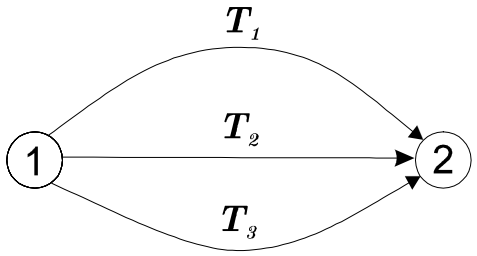

Figure 2. Parallel connection of transmittances.

The passage distribution through the mutually exclusive parallel paths is a finite mixture distribution given by the total probability formula

$$dF(t) = \sum_{i=1}^{3} p_i dF_i(t), \tag{13}$$

with transmittance

$$T(s) = \int_0^\infty e^{st} dF(t) = \sum_{i=1}^{3} p_i M_i(s) = \sum_{i=1}^{3} T_i(s) \tag{14}$$

summed over distinct paths. This is also the proper computation for the superpositioning of current flow in the parallel connection of the electrical network. If (13) is divided by dt, the resulting equation is Kirchhoff's Law for the flow of currents through a parallel connection and (14) is its transform version. The main difference, however, is that current flows through all three of the parallel paths all of the time, while the patient only takes one of the paths during one specific time increment.

In summary, whether patient movement or the flow of charge is concerned, the same mathematical techniques and formulas apply. The mathematical theory is that used in superpositioning of linear systems and is appropriate both for multistate survival models and for electrical network models.

4 Self-consistency and electrical networks

Suppose patients are subject to a single risk factor (death) but also may be independently right censored or left truncated. Section 4.2 devises a theoretical electrical network in which the flow of charge to sink describes the probability flow of such a patient to the death state while also accommodating for both censoring and truncation. In determining the impulse response at the death node, the distribution function $F^0(x)$ for patient lifetime is shown to satisfy a self-consistency relationship like that of Efron (1967) but for the population distributions that are involved.

When censored data subject to left-truncation are observed, an empirical electrical network, such as that given in Section 5, describes the flow of empirical probability reflected in the data. When determining the impulse response output at the death node, estimate $\hat{F}^0(x)$ is shown to satisfy the same self-consistency relationship as proposed in Efron (1967) and reduces to the Kaplan-Meier estimate for accommodating right censoring and left truncation. In this electrical context, the Kaplan-Meier estimate is the impulse response function reflecting the flow of empirical probability through an electrical network designed to accommodate censoring and truncation.

4.1 Censored and truncated lifetimes

Let random variable X^0 be the lifetime for a patient in the population with distribution function $F^0(x)$ and survival function $S^0(x)$. Suppose T^0 is a random truncation time that is independent of X^0 and which may also be interpreted as age upon entry into the study. Let the random censoring time Z^0 be independent of X^0 and dependent on T^0 only through the fact that the event $\{T^0 < Z^0]$ is assumed to have probability 1. If $T^0 = t$, then also assume that the conditional distribution function of $Z^0|T^0 = t$ is the distribution function G^0 restricted to (t, ∞). For an untruncated

patient who enters into the study, the truncation time T^0 is observed along with $\min(X^0, Z^0)$. This is the single event setting since one event time X^0 is considered.

Since only the smaller of X^0 and Z^0 is observed when $T^0 < \min(X^0, Z^0)$, it is convenient to define the following competitive variables and their conditional distributions:

$$T \stackrel{d}{=} T^0 \,|\, \{T^0 < X^0\} \sim E\,(t)$$
$$X \stackrel{d}{=} X^0 \,|\, \{T^0 < X^0 < Z^0\} \sim F\,(x)$$
$$Z \stackrel{d}{=} Z^0 \,|\, \{T^0 < Z^0 < X^0\} \sim G\,(z)$$

with $p_1 = P\{X^0 < Z^0 \,|\, T^0 < X^0\}$ and $p_0 = 1 - p_1$. The three random variables T, X, and Z represent competitive values for truncation time, lifetime, and censoring time respectively, and the probability p_1 is also competitive. All three distributions and p_1 are estimable from observed data. The support for all random variables is assumed to be $(0, \infty)$.

4.2 Semi-Markov systems

The lifetime of a random patient that may be right-censored and left-truncated is shown in the semi-Markov flowgraph of Figure 3. A patient is "born" into node B at time 0. The transmittance input to node B takes the value 1 and is the Laplace-Stieltjes transform for a Dirac function input at time 0. An untruncated patient enters the study in the upper portion of the flowgraph at time $T \in [t, t + dt)$ through state 1_t where t indexes one amongst a continuum of truncation-time states $\{1_t : t > 0\}$ as indicated by the triples of vertical dots. An observed lifetime occurs when the patient passes directly from $1_t \rightarrow$ D with D as the absorbing "death" state. A right-censored patient passes to state R_z amongst the continuum of right-censored states $\{\mathrm{R}_z : z > t\}$ where z is the absolute time of censoring. After censoring, the

patient's subsequent unobserved lifetime depends on z as indicated by the transition from $\mathrm{R}_z \rightarrow \mathrm{D}$. The unobserved direct transition $\mathrm{B} \rightarrow \mathrm{D}$, indicated at the bottom of the flowgraph, is the transmittance to death for a truncated patient. All transition times in the semi-Markov flowgraph are observed except for passages from $\mathrm{B} \rightarrow \mathrm{D}$ and $\mathrm{R}_z \rightarrow \mathrm{D}$.

Each pathway in the flowgraph is labelled with its transmittance. For example, the transition $\mathrm{B} \rightarrow 1_t$ occurs in time t hence the MGF is e^{st} with probability

$$dL(t) = P\{T^0 \in [t, t+dt), T^0 < X^0\} = \tau dE(t), \tag{15}$$

where $\tau = P\{T^0 < X^0\} = \int_0^\infty dL(t)$. With lifetime y, the transmittance $1_t \rightarrow \mathrm{D}$ with incremental transition time $y - t$ is

$$M_t(s) = \int_t^\infty e^{s(y-t)} dB_t(y), \tag{16}$$

where $dB_t(y)$ is the probability the patient has lifetime y observed after entering the study at time t, or

$$dB_t(y) = P\{X^0 \in [y, y+dy), Z^0 > y \,|\, T^0 \in [t, t+dt), T^0 < X^0\}. \tag{17}$$

A patient who is censored at time $z > t$ makes transition $1_t \rightarrow \mathrm{R}_z$ in time $z - t$ with probability

$$dQ_t(z) = P\{Z^0 \in [z, z+dz), X^0 > z \,|\, T^0 \in [t, t+dt), T^0 < X^0\}, \tag{18}$$

hence the transmittance $e^{s(z-t)} dQ_t(z)$. The fact that this transmittance depends on the destination state R_z is one reason why the flowgraph is semi-Markov.

Two of the transmittances $N(s, z)$ and $\Upsilon(s)$ are associated with transition times that are not directly observable. These transmittances are reexpressed in Lemma 1 in terms of quantities that are estimable as a result of the independent censoring and truncation assumptions.

Lemma 1 *Suppose that E^0, F^0, and G^0 have no common jump points so that all Riemann-Stieltjes integrals given below are defined. The transmittances that correspond to unobserved transition times are estimable though the following relationships.*

$$N(s,z) = \frac{e^{-sz}}{S^0(z)} \int_z^\infty e^{sy} dF^0(y) \tag{19}$$

$$\Upsilon(s) = \int_0^\infty e^{sy} \left\{ 1 - \tau \int_0^y \frac{dE(t)}{S^0(t)} \right\} dF^0(y). \tag{20}$$

Proof: For (19),

$$N(s,z) = \int_z^\infty e^{s(y-z)} dH_z(y), \tag{21}$$

where, for $y > z > t$,

$$\begin{aligned} dH_z(y) &= P\{X^0 \in [y, y+dy) \,|\, Z^0 \in [z, z+dz), Z^0 < X^0\} \\ &= dF^0(y)/S^0(z), \end{aligned} \tag{22}$$

upon using independent censoring. Substitution of (22) into (21) leads to (19).

Derivation of (20), requires first relating $dE(t)$ to $dE^0(t)$ as

$$\begin{aligned} dE(t) &= P\{T^0 \in [t, t+dt) \,|\, T^0 < X^0\} \\ &= P\{T^0 \in [t, t+dt), t < X^0\}/\tau \\ &= dE^0(t) S^0(t)/\tau. \end{aligned} \tag{23}$$

Thus

$$\begin{aligned} \Upsilon(s) &= \int_0^\infty e^{sy} P\left\{X^0 \in (y, y+dy), X^0 < T^0\right\} \\ &= \int_0^\infty e^{sy} \{1 - E^0(y)\} dF^0(y) \\ &= \int_0^\infty e^{sy} \left\{ 1 - \int_0^y dE^0(t) \right\} dF^0(y), \end{aligned}$$

which, upon using (23), gives (20). □

4.3 Self-consistency of F^0

The semi-Markov flowgraph in Figure 3 describes the flow of probability for the patient and so the impulse response function at node D must be the density of X^0 or $dF^0(x)/dx$. A plot of current $dF^0(x)/dx$ versus x would be the output seen on an oscilloscope were it to be attached to D. As an electrical network, the Laplace-Stieltjes transform of $F^0(x)$ is the transfer function of the network and is determined by summing all parallel transmittances from B $\rightarrow$ D. This gives

$$\int_0^\infty e^{sy} dF^0(y) = \Upsilon(s) + \Delta(s) + \Xi(s) \tag{24}$$

and represents a total probability transmittance summation for a virtual patient who may be truncated $\Upsilon(s)$, may have an observed lifetime $\Delta(s)$, or may be censored $\Xi(s)$. Here,

$$\Delta(s) = \int_{t=0}^{t=\infty} e^{st} dL(t) M_t(s) \tag{25}$$

$$\Xi(s) = \int_{t=0}^{t=\infty} e^{st} dL(t) \left\{ \int_{z=t}^{z=\infty} e^{s(z-t)} dQ_t(z) N(s,z) \right\}. \tag{26}$$

The Stieltjes integrals in (24), (25) and (26) exist so long as the distributions have no common jump points. Expressions (24), (25), and (26) may also be derived from first principles without the flowgraph presentation, however, the use of flowgraphs and the superpositioning of parallel transmittances add clarity and emphasize the electrical network analogy.

Expression (24) is actually a self-consistency equation for $F^0(x)$ expressed in the transform domain that can be solved for the value of $F^0(x)$. To do this, the transmittances $\Upsilon(s)$, $\Delta(s)$, and $\Xi(s)$ must first be expressed in terms of F^0 and the estimable distribution functions F, G, and H. Equation (20) expresses $\Upsilon(s)$ in terms of $F^0(x)$ and is substituted into (24). Expressions for $\Delta(s)$ and $\Xi(s)$ in terms of $F^0(x)$ and

$F(x)$ are

$$\Delta(s) = \tau p_1 \int_0^\infty e^{sy} dF(y) \tag{27}$$

$$\Xi(s) = \tau p_0 \int_0^\infty e^{sy} \left\{ \int_0^y \frac{dG(z)}{S^0(z)} \right\} dF^0(y), \tag{28}$$

as shown in the Appendix. When substituted into (24), the self-consistency equation for $F^0(x)$ becomes

$$\begin{aligned} \int_0^\infty e^{sy} dF^0(y) = & \int_0^\infty e^{sy} \left\{ 1 - \tau \int_0^y \frac{dE(t)}{S^0(t)} \right\} dF^0(y) \\ & + \tau p_1 \int_0^\infty e^{sy} dF(y) + \tau p_0 \int_0^\infty e^{sy} \left\{ \int_0^y \frac{dG(z)}{S^0(z)} \right\} dF^0(y). \end{aligned} \tag{29}$$

Inverting the transforms leads to the unique solution for $dF^0(x)$ as

$$\begin{aligned} dF^0(y) = & \left\{ 1 - \tau \int_0^y \frac{dE(t)}{S^0(t)} \right\} dF^0(y) \\ & + \tau p_1 dF(y) + \tau p_0 \left\{ \int_0^y \frac{dG(z)}{S^0(z)} \right\} dF^0(y). \end{aligned} \tag{30}$$

Cancelling $dF^0(y)$ on both sides, as well as the common factor τ that remains, leads to

$$dF^0(y) = \left\{ \int_0^y \frac{dE(t)}{S^0(t)} - p_0 \int_0^y \frac{dG(z)}{S^0(z)} \right\}^{-1} p_1 dF(y) \tag{31}$$

where the term in curly braces is positive. This is the population version of the self-consistent equation originally introduced by Efron (1967) for the estimation of $F^0(x)$ without truncation. If there is no truncation, then $dE(0) = 1$ so $\int_0^y dE(t)/S^0(t) = 1$ and (31) gives Efron's result for the population distribution.

Simple computations show that the right side of (31) is $dF^0(y)$. Expression (23) leads to

$$\int_0^y \frac{dE(t)}{S^0(t)} = \frac{E^0(y)}{\tau}.$$

The use of Bayes theorem on $dG(z)$ in the second term of (31) leads to

$$p_0 \int_0^y \frac{dG(z)}{S^0(z)} = p_0 \int_0^y \frac{P\left\{Z^0 \in [z, z+dz) \,|\, Z^0 < X^0, T^0 < X^0\right\}}{S^0(z)}$$
$$= p_0 \int_0^y \frac{dG^0(z)S^0(z)}{p_0 \tau S^0(z)} = \frac{G^0(y)}{\tau}.$$

Since $\{Z^0 \leq y\} \subseteq \{T^0 \leq y\}$, the term in curly braces in (31) is

$$\frac{E^0(y)}{\tau} - \frac{G^0(y)}{\tau} = \frac{P(T^0 \leq y < Z^0)}{P(T^0 \leq X^0)}.$$

The right side of (31) is now

$$\left\{\frac{P(T^0 \leq y < Z^0)}{P(T^0 \leq X^0)}\right\}^{-1} p_1 P\{X^0 \in [y, y+dy) | T^0 < X^0 < Z^0\} = dF^0(y)$$

when Bayes theorem is used on the last probability for $dF(y)$.

5 Kaplan-Meier and the flow of empirical probability

Suppose untruncated data consist of n_1 lifetimes, observed as the pairs $\{(t_{1i}, x_i) : i = 1, \ldots, n_1\}$ where truncation time $t_{1i} < x_i$, and n_0 censored values $\{(t_{0j}, z_j) : j = 1, \ldots, n_0\}$ with $t_{0j} < z_j$. Distribution functions E, F, and G are estimated by their empirical counterparts $\hat{E}(t), \hat{F}(x)$, and $\hat{G}(z)$ based on $\{t_{1i}\} \cup \{t_{0j}\}$, $\{x_i\}$, and $\{z_j\}$ respectively while $\hat{p}_1 = n_1/n_\cdot$ with $n_\cdot = n_0 + n_1$.

Figure 4 shows a semi-Markov flowgraph that is an empirical version of the graph in Figure 3. Each patient contributes a separate path from B $\rightarrow$ D with weight $n_\cdot^{-1}\tau$. If estimate $\hat{F}^0(y) = 1 - \hat{S}^0(t)$ is assumed to exist, then unobserved branches $\mathrm{R}_{z_j} \rightarrow$ D and B $\rightarrow$ D direct have empirical transmittances

$$\hat{N}(s, z_j) = \frac{e^{-sz_j}}{\hat{S}^0(z_j)} \int_{z_j}^{\infty} e^{sy} d\hat{F}^0(y)$$
$$\hat{\Upsilon}(s) = \int_0^{\infty} e^{sy} \left\{1 - \int_0^y \frac{\tau d\hat{E}(t)}{\hat{S}^0(t)}\right\} d\hat{F}^0(y)$$

obtained by using the estimable expressions in Lemma 1. Superpositioning by summing over parallel empirical transmittances, gives a transfer function for the output at node D as

$$\int_0^\infty e^{sy} d\hat{F}^0(y) = \hat{\Upsilon}(s) + \frac{\tau}{n_\cdot} \sum_{i=1}^{n_1} e^{st_{1i}} e^{s(x_i - t_{1i})} + \frac{\tau}{n_\cdot} \sum_{j=1}^{n_0} e^{st_{0j}} e^{s(z_j - t_{0j})} \hat{N}(s, z_j). \tag{32}$$

The second term is $\tau n_1/n_\cdot \int_0^\infty e^{sy} d\hat{F}(y)$ while the last term is

$$\frac{\tau}{n_\cdot} \int_0^\infty \frac{1}{\hat{S}^0(z)} \left\{ \int_z^\infty e^{sy} d\hat{F}^0(y) \right\} n_0 d\hat{G}(z) = \frac{\tau n_0}{n_\cdot} \int_0^\infty e^{sy} \left\{ \int_0^y \frac{d\hat{G}(z)}{\hat{S}^0(z)} \right\} d\hat{F}^0(y).$$

Inverting the Laplace-Stieltjes transforms in (32) and rearranging terms leads to

$$\hat{C}(y) d\hat{F}^0(y) := \left\{ \int_0^y \frac{d\hat{E}(t)}{\hat{S}^0(t)} - \hat{p}_0 \int_0^y \frac{d\hat{G}(z)}{\hat{S}^0(z)} \right\} d\hat{F}^0(y) = \hat{p}_1 d\hat{F}(y) \tag{33}$$

as the defining equation for self-consistency. The solution to (33) is now summarized.

Theorem 2 *Let $x_* = \min\{x_i\}$ and $x^* = \max(\{x_i\}, \{z_j\})$. A unique self-consistent solution exists to (33) over (x_*, x^*) which is the Kaplan-Meier estimator. As a consequence, the Kaplan-Meier mass points at $\{x_i\}$ comprise the discrete impulse response function at node D for the flow of empirical probability through the network in Figure 4. These results require the following conditions: (i) $\hat{E}(t)$, $\hat{F}(t)$, and $\hat{G}(t)$ have no common jump points; (ii) without any loss in generality, censored values less than x_* have already been deleted as uninformative; (iii) if $\mathcal{N}_t$ is the number of patients at risk at time t, then $\mathcal{N}_t > 0$ for all $t \in (x_*, x^*)$. Assumption (iii) assures that the solution is unique.*

Proof. Since $d\hat{F}(y) = 0$ for $y \notin \{x_i\}$, the support for $d\hat{F}^0(y)$ can only be $\{x_i\}$ and also regions in (x_*, x^*) for which $\hat{C}(y) = 0$. The latter possibility will be eliminated with the assumption that $\mathcal{N}_t > 0$ for all $t \in (x_*, x^*)$. For the sake of argument, suppose that $x_1 < x_2 < \cdots < x_{n_1}$. For any t between $\min[\{t_{1i}\}, \{t_{0j}\}]$, and x_1, $\hat{C}(t) > 0$ and

and $d\hat{F}(t) = 0$ imply $d\hat{F}^0(t) = 0$ so that $\hat{S}^0(t) = 1$ over this range. At $x_1 = x_*$, (33) is

$$\hat{C}(x_1)d\hat{F}^0(x_1) = \left(\frac{\mathcal{N}_1}{n.} - 0\right) d\hat{F}^0(x_1) = \frac{1}{n.} \tag{34}$$

where $\mathcal{N}_1$ denotes the number truncated before x_1 and therefore at risk, while 0 is the number censored before x_1. The solution to (34) is $d\hat{F}^0(x_1) = 1/\mathcal{N}_1$ so that $\hat{S}(x_1) = 1 - 1/\mathcal{N}_1$.

More generally, the following recursions

$$\frac{1}{d\hat{F}^0(x_l)} = \frac{\mathcal{N}_l}{\hat{S}^0(x_{l-1})} \qquad \text{and} \qquad \hat{S}^0(x_l) = \hat{S}^0(x_{l-1})\left(1 - \frac{1}{\mathcal{N}_l}\right) \tag{35}$$

are shown to hold for $l = 1, \ldots, n_1$ which are those for the Kaplan-Meier estimator. The $l = 1$ case above has been shown to hold with $x_0 = x_1^-$.

The proof of (35) proceeds by using induction wherein recursion l in (35) is shown to imply recursion $l + 1$. Suppose that $\Delta T_{l+1} = n.\{\hat{E}(x_{l+1}) - \hat{E}(x_l)\}$ and $\Delta R_{l+1} = n.\hat{p}_0\{\hat{G}(x_{l+1}) - \hat{G}(x_l)\}$ count the number of truncation and right censoring times within (x_l, x_{l+1}). From (33),

$$\begin{aligned}\frac{1}{d\hat{F}^0(x_{l+1})} = \frac{1}{n.\hat{C}(x_{l+1})} &= \sum_{k=0}^{1} \sum_{\{j\,:\,t_{kj} < x_{l+1}\}} \frac{1}{\hat{S}^0(t_{kj})} - \sum_{\{j\,:\,z_j < x_{l+1}\}} \frac{1}{\hat{S}^0(z_j)} \\ &= \frac{1}{d\hat{F}^0(x_l)} + \frac{\Delta T_{l+1}}{\hat{S}^0(x_l)} - \frac{\Delta R_{l+1}}{\hat{S}^0(x_l)},\end{aligned}$$

where the constant value $\hat{S}^0(t) \equiv \hat{S}^0(x_l)$ over $t \in [x_l, x_{l+1})$ has been used. From (35), this is

$$\begin{aligned}\frac{1}{d\hat{F}^0(x_{l+1})} &= \frac{\mathcal{N}_l}{\hat{S}^0(x_{l-1})} + \frac{\Delta T_{l+1} - \Delta R_{l+1}}{\hat{S}^0(x_{l-1})\,(1 - 1/\mathcal{N}_l)} \\ &= \frac{\mathcal{N}_l}{\hat{S}^0(x_{l-1})\,(\mathcal{N}_l - 1)}(\mathcal{N}_l - 1 + \Delta T_{l+1} - \Delta R_{l+1}) \\ &= \frac{\mathcal{N}_{l+1}}{\hat{S}^0(x_l)},\end{aligned} \tag{36}$$

upon using the recursion for $\hat{S}^0(x_l)$ in (35). Using $\hat{S}^0(x_{l+1}) = \hat{S}^0(x_l) - d\hat{F}^0(x_{l+1})$ and (36) gives the remaining recursion. □

5.1 Relevant Literature

Semi-Markov modeling has been used to explain right censoring (but without truncation) in Lagakos, Sommer, & Zelen (1978) and later Anderson *et al.* (1993, Ex. III2.8). Their models record passage time up to $\min(X^0, Z^0)$ but not afterwards and alternatively represent censoring as an absorbing node that competes with the death node. Their semi-Markov competing risks models are different and only show branches of the model that are observable. By contrast, our semi-Markov models take account of the unobservable transitions that include life transitions after censoring, $\mathrm{R}_z \to \mathrm{D}$, and truncation without observation, $\mathrm{B} \to \mathrm{D}$. By accounting for all transitions leading to state D, the impulse response at state D is the density for X_0. This cannot happen in these other model settings where censoring and death are competing as absorbing states in the model. Of course these other models were developed to address different concerns: the determination of a nonparametric maximum likelihood estimator for F^0. Perhaps what is most interesting is that both models lead to the same Kaplan-Meier estimator.

A Markov model that allows for random left truncation was also given in Anderson *et al.* (1993, Ex. III3.3). This model shows the parallel connection of untruncated and truncated paths but otherwise has not been expanded to also allow for right censoring as in Figure 3.

5.2 Irregularities and examples

Uniqueness of the solution to the self-consistent equations in (33) may be lost if no patients are at risk during a portion of the informative time span, i.e., there is a $t^0 \in (x_*, x^*)$ for which $\mathcal{N}_{t^0} = 0$. Furthermore, in this instance $\hat{S}^0$ does not have to place all mass on $\{x_i\}$ but rather can place non-zero mass on an interval that

contains t^0 since $\hat{C}(t^0) = 0$. Let t^* be the next observed time point above t^0, which is necessarily a truncation time, and let z_* be the next time point below t^0 which is necessarily a lifetime or censoring time. If z_* is a censoring time (lifetime), then interval (z_*, t^*) can (cannot) hold non-zero mass in the solution to the self-consistency equations. Support intervals such as (z_*, t^*) were first noted by Frydman (1994) as additional sites capable of holding mass for the nonparametric maximum likelihood estimate when there is truncation. Such sites were not mentioned in Turnbull's (1976) original account dealing with general interval censoring and truncation. These points are illustrated using two simple examples.

Example 1. Consider the ordered data

$$t_{x_1} < t_{x_2} < x_1 < t_{z_1} < z_1 < x_2 < t_{x_3} < x_3$$

in which t_{x_1} is the truncation time for x_1, etc. At x_2, the two patients entered into the study are no longer at risk but the third patient has not yet entered. The example violates the conditions of Theorem 2 since there are no patients at risk during the interval (x_2, t_{x_3}). The self-consistent solution places mass $1/2$ on x_1 and $1/2$ on x_2. This leads to $\hat{S}^0(x_2) = 0$ and $\hat{C}(x_2) = 2$ from which the value for

$$\hat{C}(t_{x_3}) = \hat{C}(x_2) + \frac{1}{\hat{S}^0(t_{x_3})}$$

is undefined due to division by $\hat{S}^0(t_{x_3}) = 0$.

In computing the nonparametric maximum likelihood estimate, the support set from Turnbull (1976) is x_i with probability s_i for $i = 1, 2, 3$. The respective likelihood terms contributed by x_1, z_1, x_2, and x_3 are

$$L = s_1(s_2 + s_3)\frac{s_2}{s_2 + s_3}\frac{s_3}{s_3} = s_1 s_2$$

and the maximum likelihood estimate agrees with the self-consistent estimate with $\hat{s}_1 = \hat{s}_2 = 1/2$.

Example 2. Interchange z_1 and x_2 from Example 1 to get ordered data

$$t_{x_1} < t_{x_2} < x_1 < t_{z_1} < x_2 < z_1 < t_{x_3} < x_3.$$

Again there are no patients at risk during (z_1, t_{x_3}) so Theorem 2 is violated. The self-consistent solution places mass $1/2$ on x_1, $1/4$ on x_2, mass p in (z_1, t_{x_3}), and mass $1/4 - p$ at x_3 for any $p \in [0, 1/4)$. In the self-consistent solution,

$$\hat{C}(z_1) = 4 - \frac{1}{\hat{S}^0(z_1)} = 4 - 4 = 0.$$

By allowing arbitrary mass p in (z_1, t_{x_3}), then

$$\hat{C}(t_{x_3}) = \hat{C}(z_1) + \frac{1}{\hat{S}^0(t_{x_3})} = \frac{1}{1/4 - p} > 0$$

for any $p \in [0, 1/4)$ which leaves $d\hat{F}^0(x_3) = 1/4 - p$.

The support set for the maximum likelihood estimate is determined from Frydman (1994) as x_1, x_2, (z_1, t_{x_3}), and x_3 with probabilities $s_1, \ldots, s_4$. The nonparametric likelihood is

$$L = s_1 s_2 \frac{s_3 + s_4}{s_2 + s_3 + s_4} = s_1 s_2 \frac{1 - s_1 - s_2}{1 - s_1}$$

which attains the same collection of maxima as the self-consistent solution.

5.3 Proportional hazards extensions in single event settings

Suppose data consist of n patients with responses $(t_i, x_i, \delta_i, u_i)$ for $i = 1, \ldots, n$ where the respective values are truncation time, lifetime/censoring time, indicator of lifetime response, and covariate vector. For notational convenience suppose that $x_1 < \cdots < x_n$ and that there are no ties. In the context of the proportional hazards model, patient i with lifetime X_{θ_i} has survival function $P(X_{\theta_i} > t) = S_0(t)^{\theta_i}$ with $\theta_i = \exp(\beta^T u_i)$. This patient's survival function is the same as that for θ_i independent virtual baseline patients, hence, in electrical terms, this patient has a charge of θ_i

coulombs. Equivalently, the hazard for X_{θ_i} is the sum of the hazards of these θ_i virtual baseline patients.

By working with a total charge of $\theta_\cdot = \sum_{i=1}^n \theta_i$ coulombs or virtual patients rather than n heterogeneous patients, partial likelihood and baseline survival $S_0(t)$ estimates take on new interpretations. For an assumed β, the partial likelihood under both right censoring and left truncation is

$$L_p(\beta) = \prod_{i=1}^{n} \left(\frac{\theta_i}{\sum_{j \in R_i} \theta_j} \right)^{\delta_i}, \tag{37}$$

where R_i is the risk set at time x_i so that $j \in R_i$ whenever $t_j < x_i < x_j$. At death point x_i, the hazard contribution to partial likelihood is the proportion of charge or the fraction of virtual patients that die at that time point. It is the proportional hazards model assumption itself that allows a risk set of R_i heterogenous patients to be summarized by putting $\sum_{j \in R_i} \theta_j$ coulombs of charge at risk at time x_i amongst which θ_i coulombs are dissipated.

The Breslow or Nelson-Aalen estimate of baseline survival for an assumed value of β is simply

$$\hat{S}_0(t) = \prod_{\{i \,:\, x_i \le t\}} \left(1 - \frac{1}{\sum_{j \in R_i} \theta_j} \right)^{\delta_i}. \tag{38}$$

Estimate (38) is meaningful under the electrical analog because $1 - 1/\sum_{j \in R_i} \theta_j$ is the probability that 1 coulomb or a single baseline patient survives lifetime point x_i when $\sum_{j \in R_i} \theta_j$ coulombs or virtual baseline patients are subject to risk at time x_i^-. Thus (38) adjusts the observed probabilities pertaining to θ_i-patients so they are relevant to a single baseline patient.

The Kalbfleisch & Prentice estimator (2002, eqn. 4.36)

$$\hat{S}_0^{KP}(t) = \prod_{\{i \,:\, x_i \le t\}} \left(1 - \frac{\theta_i}{\sum_{j \in R_i} \theta_j} \right)^{\delta_i/\theta_i} \tag{39}$$

is equally simple to motivate. The probability that θ_i coulombs survive time x_i is $1 - \theta_i / \sum_{j \in R_i} \theta_j$ when computed from amongst $\sum_{j \in R_i} \theta_j$ coulombs at risk. However, this is the survival probability $S_0(t)^{\theta_i}$ for X_{θ_i} not X_1. To have this apply to a baseline patient, the probability is raised to the $1/\theta_i$ power as shown in (39).

The electrical analogy equates the risk of a θ_i-patient with θ_i independent baseline patients. From this perspective, very simple and intuitive interpretations result for these three commonly used estimators.

6 Competing Risks

In the classical competing risks setting, there are multiple event times $X_1^0, \ldots, X_K^0$ with distribution $F^0(x_1, \ldots, x_K)$ that compete with independent censoring time Z^0 and independent left truncation time T^0. The value and index for $M^0 = \min\{X_k^0\}$ are observed if the events are untruncated, $T^0 < M^0$, and uncensored, $M^0 < Z^0$. The aim is to estimate the collection of subdistributions

$$F_k^0(x) = P(X_k = M^0 \leq x) \qquad k = 1, \ldots, K$$

associated with F^0 from competitive data that are subject to right censoring and left truncation. The distributional structure of the censoring and truncation variables supposes $T^0 \sim E^0$ and $Z^0 | T^0 = t \sim G^0$ restricted to (t, ∞) with T^0 and Z^0 independent of $X_1^0, \ldots, X_K^0$. The survival function of M^0 is $S_+^0(t) = 1 - F_+^0(t)$ where $F_+^0(t) = \sum_{k=1}^K F_k^0(x)$.

The data are observed to come from the competitive distributions

$$\begin{aligned}
T &\stackrel{d}{=} T^0 \,|\, \{T^0 < X^0\} \sim E\,(x) \\
X_k &\stackrel{d}{=} X_k^0 \,|\, \{T^0 < X_k^0 = M^0 < Z^0\} \sim F_k\,(x) \\
Z &\stackrel{d}{=} Z^0 \,|\, \{T^0 < Z^0 < M^0\} \sim G\,(z)\,.
\end{aligned}$$

for $k = 1, \ldots, K$. The competitive probabilities $p_k = P(X_k^0 = M^0 < Z^0)$ for $k \geq 1$ and $p_0 = P(Z^0 < M^0)$ are estimable from the data and add to 1.

6.1 Population semi-Markov flowgraph

Figure 5 shows the competing risk flowgraph when there are $K = 2$ possible events that are subject to random right censoring and left truncation. The transmittances along with estimable expressions, determined from arguments similar to those in section 4.2, are $dL(t) = \tau dE(t)$ where $\tau = P(T^0 < M^0)$ and $M_{kt}(s) = \int_t^\infty e^{s(y-t)} dB_k(y)$ where

$$
\begin{aligned}
dB_k(y) &= P\left\{X_k^0 = M^0 \in [y, y+dy), M^0 < Z^0 \,|\, T^0 \in [t, t+dt), T^0 < M^0\right\} \\
dQ_t(z) &= P\left\{Z^0 \in [z, z+dz), Z^0 < M^0 \,|\, T^0 \in [t, t+dt), T^0 < M^0\right\} \\
N_k(s,z) &= \frac{e^{-sz}}{S_+^0(z)} \int_z^\infty e^{sy} dF_k^0(y) \\
\Upsilon_k(s) &= \int_0^\infty e^{sy} \left\{1 - \tau \int_0^y \frac{dE(t)}{S_+^0(t)}\right\} dF_k(y).
\end{aligned}
$$

The sum over all paths from B to D_k, within which lifetimes of type k are observed, gives transmittance

$$\Delta_k(s) = \tau p_k \int_0^\infty e^{sy} dF_k(y).$$

Summing from B to D_k but passing through censored states $\{\mathrm{R}_z\}$ gives

$$\Xi_k(s) = \tau p_0 \int_0^\infty e^{sy} \left\{\int_0^y \frac{dG(z)}{S_+^0(z)}\right\} dF_k^0(y).$$

The Laplace-Stieltjes transform for subdistribution F_k^0 is the sum over all transmittances from B to D_k or $\Upsilon_k(s) + \Delta_k(s) + \Xi_k(s)$. Transform inversion can be shown to lead to

$$dF_k^0(y) = \left\{\int_0^y \frac{dE(t)}{S_+^0(t)} - p_0 \int_0^y \frac{dG(z)}{S_+^0(z)}\right\}^{-1} p_k \, dF_k(y) \tag{40}$$

for $k = 1, \ldots, K$.

6.2 Empirical semi-Markov flowgraph

The data consist of n_k observed events of type k and n_0 censored with total sample size $n_\cdot = n_0 + \sum_{k=1}^{K} n_k$. Suppose x_{ki} is the ith observed event time of type k with associated truncation time $t_{ki} < x_{ki}$. The times $\{t_{0i} : i = 1, \ldots, n_0\}$ are the truncation times for the censored data. Truncation distribution E is estimated using $\hat{E}$, the empirical distribution of $\{t_{kj} : k = 0, \ldots, K; j = 1, \ldots, n_k\}$ while $\hat{F}_k$ and $\hat{G}$ are the empirical distributions of $\{x_{ki} : i = 1, \ldots, n_k\}$ and censoring times $\{z_j : j = 1, \ldots, n_0\}$, respectively. Event probability estimates are $\hat{p}_k = n_k/n_\cdot$.

Figure 6 shows the empirical competing risk flowgraph with $K = 2$. Summing over all paths from B to D_k leads to the self-consistency equations

$$\hat{C}(y)d\hat{F}_k^0(y) = \hat{p}_k d\hat{F}_k(y) = 1/n_\cdot \qquad k = 1, \ldots, K \tag{41}$$

where

$$\hat{C}(y) = \int_0^y \frac{d\hat{E}(t)}{\hat{S}_+^0(t)} - \hat{p}_0 \int_0^y \frac{d\hat{G}(z)}{\hat{S}_+^0(z)}.$$

The solution to (41) is now summarized.

Theorem 3 *Let $x_* = \min_{k,i}\{x_{ki}\}$ and $x^* = \max_{k,i,j}(\{x_{ki}\}, \{z_j\})$. Subject to conditions (i)-(iii) below, a unique self-consistent solution exists to the K equations in (41) over (x_*, x^*). The solution has subdistribution mass functions given as the Kaplan-Meier weights for this competing risks setting. These Kaplan-Meier weights are the discrete impulse response functions at their respective nodes $\{D_k\}$ that result from the flow of empirical probability through the network in Figure 6 (adjusted if $K > 2$). These results require the following conditions: (i) $\hat{E}(t)$, $\{\hat{F}_k(t)\}$, and $\hat{G}(t)$ have no common jump points; (ii) without any loss in generality, censored values less than x_* have already been deleted as uninformative; (iii) if $\mathcal{N}_t$ is the number of patients at risk at time t, then $\mathcal{N}_t > 0$ for all $t \in (x_*, x^*)$. Assumption (iii), in particular, assures that the solution is unique.*

Proof. The proof is the same inductive proof used for Theorem 2. Assumption $\mathcal{N}_t > 0$ assures that mass is only placed on event times. Order the $N = n_\cdot - n_0$ event times as $x_1 < \cdots < x_N$ and suppose the associated event types are $i(1), \ldots, i(N)$. The $l = 1$ case of induction holds and states that $d\hat{F}^0_{i(1)}(x_1) = 1/\mathcal{N}_1$ and $\hat{S}^0_+(x_1) = 1 - 1/\mathcal{N}_1$ where $\mathcal{N}_1$ is the number of patients at risk at time $x_1 = x_*$. Assuming the lth case specified as

$$\frac{1}{d\hat{F}^0_{i(l)}(x_l)} = \frac{\mathcal{N}_l}{\hat{S}^0_+(x_{l-1})} \qquad \hat{S}^0_+(x_l) = \hat{S}^0_+(x_{l-1})\,(1 - 1/\mathcal{N}_l)\,,$$

it can then be shown that the $(l+1)$st case holds. This gives an inductive proof for Kaplan-Meier weights. □

Subdistribution estimate $\hat{F}^0_k$ accumulates Kaplan-Meier probabilities $d\hat{F}^0_k$ at event times of type k where the Kaplan-Meier estimate $\hat{S}^0_+(x) = 1 - \sum_{k=1}^K \hat{F}^0_k(x)$ for the survival of M^0 has been computed by using the pooled set of event times. If x^* is a censored value, then $\hat{S}^0_+(x)$ and $\{\hat{F}^0_k(x)\}$ are indeterminate for $x > x^*$.

6.3 Proportional hazards extensions in competing risk settings

Prentice *et al.* (1978) review the options proposed by Holt (1978) for including covariates in the Cox model and offer two models for this setting. In the first, regression coefficients are cause-specific so that a patient with covariate u would have the cause-specific hazard $\theta_k \lambda_{0k}(t)$ where $\theta_k = \exp(\beta_k^T u)$ and the baseline hazard is $\lambda_{0k}(t) = dF^0_k(t)/S^0_+(t)$. For this model, the K cause-specific parameter sets $\{\theta_1, \lambda_{01}(\cdot)\}, \ldots, \{\theta_K, \lambda_{0K}(\cdot)\}$ are L-independent (Barndorff-Nielsen, 1978, §3.3) in the sense that the likelihood is completely separable into K such groups of parameters. Accordingly, cause specific baseline hazards, subdistributions, and regressions are estimated separately by using the methods of section 5.3.

Under the second model, the cause specific baseline hazards are proportional so the kth event type hazard is $\theta_k \lambda_0(t)$ with $\theta_k = \exp(\alpha_k + \beta_k^T u)$. In this model, a patient with covariate u is assumed to have $\theta_\cdot = \sum_{k=1}^K \theta_k$ coulombs of charge and can be shown to have marginal survival function $S_0(t)^{\theta_\cdot}$. In regression estimation, this leads to the partial likelihood given in Prentice *et al.* (1978, p. 547) that factors a logistic model $\{\theta_j/\theta_\cdot : j = 1, \ldots, K\}$ for the event type of each patient as well as a partial likelihood factor as in (37) that computes probabilities for deaths of patients but assuming patients have charges of $\theta_\cdot$ that represent their total risk. Correspondingly, estimates of $S_0(t)$ have support over the pooled collection of event times and can be computed as in (38) or (39) but again assuming each patient has $\theta_\cdot$ coulombs of charge.

7 Appendix

Proof of (27): Substituting (15), (16), and (17) into (25) gives

$$\begin{aligned}
\Delta(s) &= \tau \int_{t=0}^{t=\infty} e^{st} dE(t) \int_{y=t}^{y=\infty} e^{s(y-t)} dB_t(y) \\
&= \tau \int_{y=0}^{y=\infty} e^{sy} \int_{t=0}^{t=y} P\left\{X^0 \in [y, y+dy), Z^0 > y, T^0 \in [t, t+dt) \,|\, T^0 < X^0\right\} \\
&= \tau \int_0^\infty e^{sy} P\{X^0 \in [y, y+dy), Z^0 > X^0 \,|\, T^0 < X^0\} = \tau p_1 \int_0^\infty e^{sy} dF(y).
\end{aligned}$$

Proof of (28): Substituting (15) and (19) into (26) gives

$$\Xi(s) = \tau \int_{t=0}^{t=\infty} dE(t) \int_{z=t}^{z=\infty} \frac{dQ_t(z)}{S^0(z)} \int_z^\infty e^{sy} dF^0(y) = \int_0^\infty e^{sy} dB(y),$$

where

$$dB(y) = \tau dF^0(y) \int_{z=0}^{z=y} \frac{1}{S^0(z)} \int_{t=0}^{t=z} dQ_t(z) dE(t). \tag{42}$$

Substituting (18) into (42), the integral in t is

$$\begin{aligned}\int_{t=0}^{t=z} & P\{Z^0 \in [z, z+dz), X^0 > Z^0, T^0 \in [t, t+dt) \,|\, T^0 < X^0\} \\ &= P\{Z^0 \in [z, z+dz), X^0 > Z^0 \,|\, T^0 < X^0\} \\ &= P\{Z^0 \in [z, z+dz) \,|\, T^0 < Z^0 < X^0\} P\{X^0 > Z^0 \,|\, T^0 < X^0\} \\ &= dG(z) p_0. \end{aligned} \tag{43}$$

Substitution of (43) into (42) gives (28).

References

[1] Andersen, P.K., Borgan, Ø, Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

[2] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory.* New York: Wiley.

[3] Butler, R.W. (2000). Reliabilities for feedback systems and their saddlepoint approximation. *Statist. Sci.* **15**, 279-298.

[4] Butler, R.W. (2001). First passage distributions in semi-Markov processes and their saddlepoint approximation. In *Data Analysis from Statistical Foundations* Ed. E. Saleh, pp. 347-68. Huntington, NY: Nova Science Publishers.

[5] Butler, R.W. (2007). *Saddlepoint Approximations with Applications.* Cambridge: Cambridge University Press.

[6] Doyle, P.G. and Snell, J.L. (1984). *Random Walks and Electrical Networks.* Carus Mathematical Monographs 22, Mathematical Association of America.

[7] Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **4**, 831-53.

[8] Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *J. R. Statist. Soc. B* **56**, 71-4.

[9] Holt, J.D. (1978). Competing risk analyses with special reference to matched pair experiments. *Biometrika* **65**, 159-66.

[10] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data* 2nd ed. New York: Wiley.

[11] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457-81.

[12] Kelly, F.P. (1978). *Reversibility and Stochastic Networks.* New York: Wiley.

[13] Lagakos, S. W., Sommer, C. J. & Zelen, M. (1978). Semi-Markov models for partially censored data. *Biometrika* **65**, 311-7.

[14] Medhi, J. (1994). *Stochastic Processes.* New York: Wiley.

[15] Miller, R. (1981). *Survival Analysis.* New York: Wiley.

[16] Phillips, C.L. and Harbor, R.D. (1996). *Feedback Control Systems.* New Jersey: Prentice Hall.

[17] Prentice, R.L., Kalbfleisch, J.D., Peterson Jr., A.V., Flournoy, N, Farewell, V.T. & Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541-54.

[18] Pyke, R. (1961). Markov renewal processes with finitely many states. *Ann. Math. Stat.* **32**, 1243-59.

[19] Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc. B* **38**, 290-5.

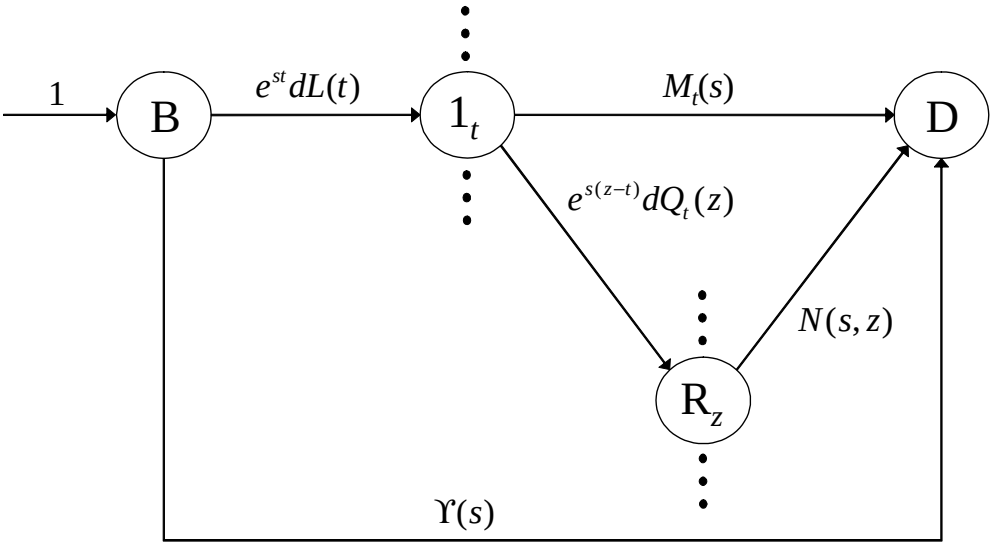

Figure 3. Semi-Markov flowgraph for a virtual patient's lifetime subject to a single risk.

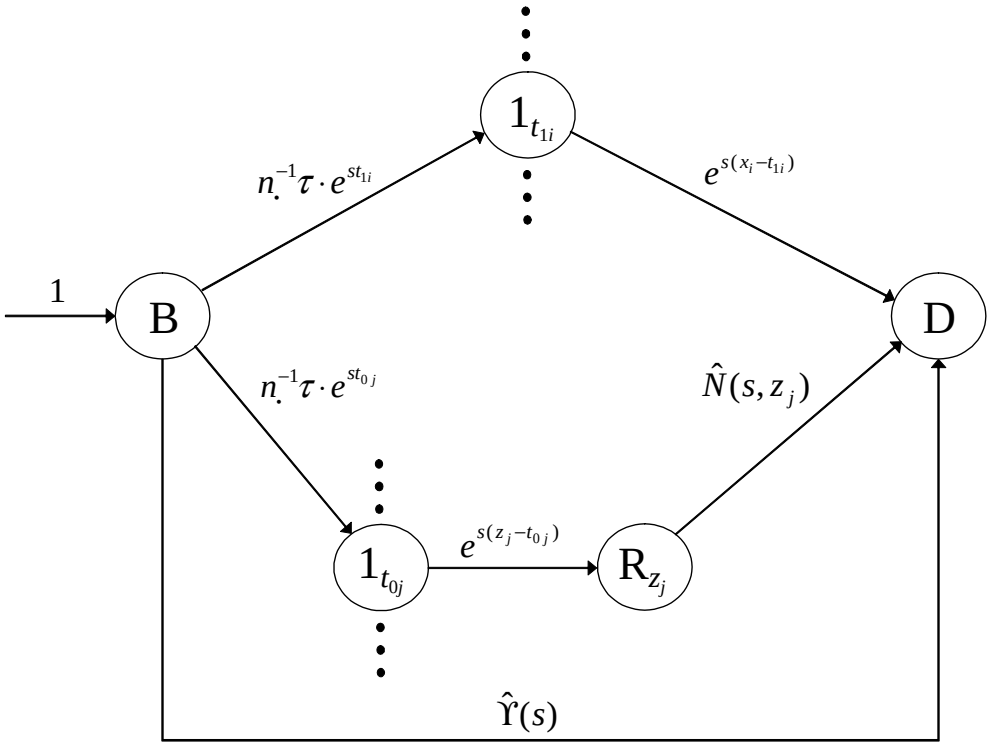

Figure 4. Empirical semi-Markov flowgraph providing an approximation for the single-risk flowgraph of Figure 5.

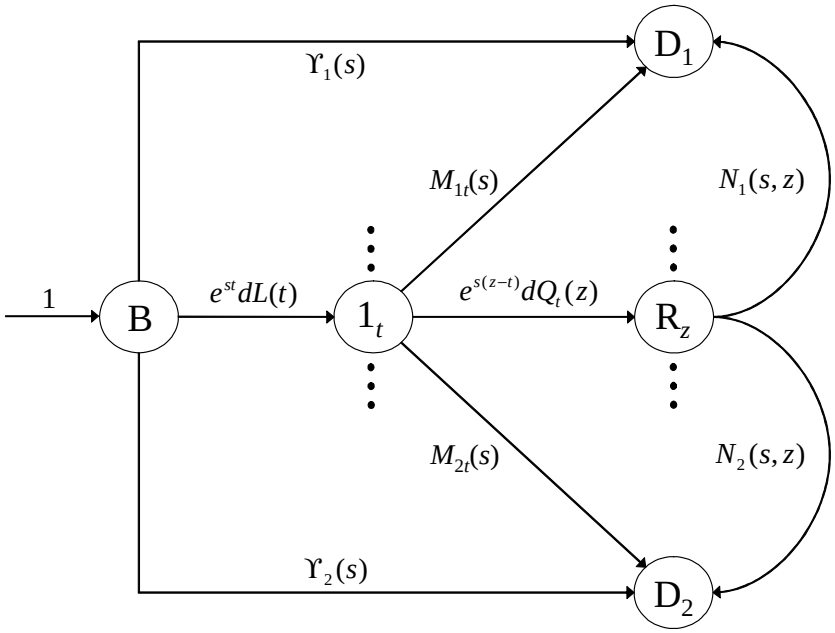

Figure 5. Semi-Markov flowgraph for a virtual patient's lifetime subject to two competing risks.

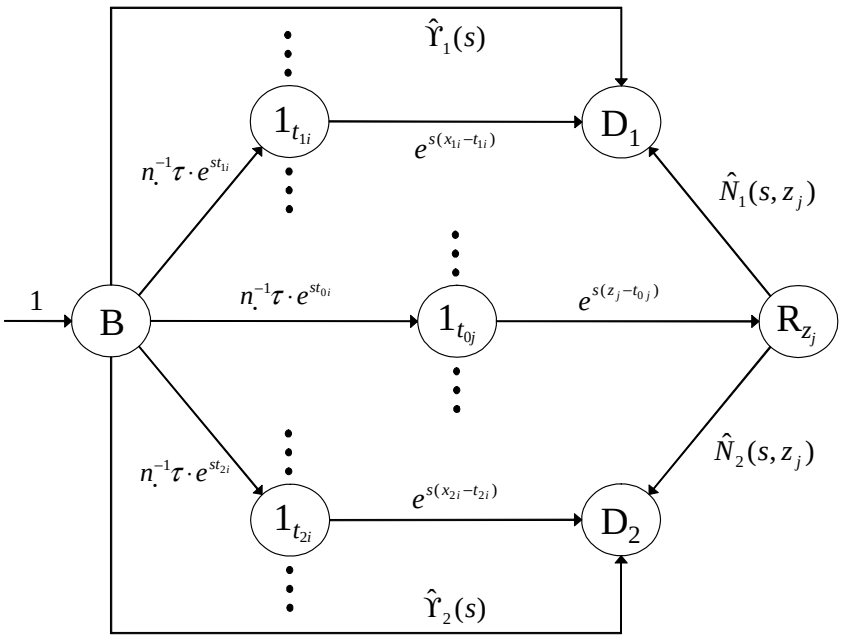

Figure 6. Empirical flowgraph providing a semi-Markov approximation for the competing-risks flowgraph in Figure 7.