Nonidentifiability in stochastic models of illness and death

(Peter Clifford*)

*Statistical Laboratory, Evans Hall, University of California, Berkeley, California 94720

Communicated by J. Neyman, January 19, 1977

ABSTRACT  The interpretation of animal survival experiments in which disease incidence is determined at death or following sacrifice is shown to involve certain ambiguities. In particular, quantities of interest such as the expected duration of life for an animal contracting a specific disease at a specific age are found to be nonidentifiable. An example is constructed in which two populations of animals will appear similar to the experimenter but in which animals contracting a particular disease in one population may have double the life expectancy of similarly afflicted animals in the other population.

In a simple survival experiment $N$ newborn animals are observed for the duration of their lives. For each animal, the age $X_i$ at death and cause $C_i$ of death are recorded, $i = 1, 2, \ldots, N$. Denoting the set of distinct causes of death by $D$ we have $X_i \geq 0$ and $C_i \in D$, $i = 1, 2, \ldots, N$. The animals will be assumed to die independently and to be members of a homogeneous population.

An attempt is often made to use such experiments for prediction. Thus, we may ask what would be the effect of "removing" (by improved health care, say) one of the causes of death. In order to make predictions a model has to be introduced. The potential survival time model has repeatedly been proposed in this context. The method is reviewed by David (1).

In this model it is supposed that for each animal and for each cause of death $d \in D$ there is an associated nonnegative random variable $Y_d$ called the potential survival time. The distribution of age at death $X$ is assumed to be the same as the distribution of $\min_{d \in D} Y_d$ and the associated cause of death is the subscript of the minimizing $Y$ variable. If the distribution of the $Y_d$'s can be determined from the data then it is argued that the distribution of lifespan when only a reduced set of causes of death $D^*$ are present will be given by $\min_{d \in D^*} Y_d$. Tsiatis (2) has shown that the $Y$'s cannot be uniquely determined from records which consist solely of age at death and cause of death. In particular it is impossible to determine whether or not the $Y$'s are independent. Prediction may thus be so ambiguous as to be purely of academic interest.

Serial sacrifice

Such difficulties seem to be known, at least in principle, to the biological community. For example, the survival and serial sacrifice experiments of Upton et al. (3) on radiation-induced cancers in mice were designed to "elucidate . . . mechanisms" of disease development. They differ from simple survival experiments in that recorded data include post mortem examinations for the presence of seven principal diseases and that a number of living animals were sacrificed and examined over a 30 month period to "provide data on time of onset and rate of development" of diseases. There are therefore $2^7$ possible disease combinations or states which could be recorded after post mortem examination. We will show that these data do permit us to test whether diseases progress independently; however, certain ambiguities remain that cannot be resolved with this type of experiment. These comments are made in relation to a certain structural model for the progress of disease, the Markov Illness Death model.

Markov Illness Death model

This model was discussed by Neyman (4) and applied by Fix and Neyman (5) to follow-up studies of cancer treatment. Chiang (6) treats the subject in great generality. A large number of parameters are involved in the model, making estimation difficult. This number may be reduced when it is assumed that all diseases are progressive, i.e., that recovery from a disease is impossible. To facilitate estimation we will make this assumption.

The model is essentially a compartment model, where disease states correspond to compartments. As an example the network with three diseases $a, b, c$ and their combinations is depicted in Fig. 1. $H$ denotes the healthy state. Animals are assumed to start in the healthy state. The arrows on the paths between states correspond to the direction of transitions, and the symbols to the intensity of transition. These parameters will in general be functions of time or age. This complication is introduced not only because it is biologically realistic but also because this is the class of alternatives against which the null hypothesis of constant intensities must be critically reviewed. The intensities have the usual interpretation, e.g., the probability that an animal with disease $a$ at age $t$ contracts disease $b$ in time $(t, t + \tau)$ is $\mu_{ab}(t)\tau + o(\tau)$.

From each state $a$ there is also a death intensity $\mu_a(t)$ so that the probability that an animal in state $a$ at age $t$ dies in $(t, t + \tau)$ is $\mu_a(t)\tau + o(\tau)$. The process is clearly Markov because only the current state of the animal influences its future. If the parameters of the model can be estimated, tests of hypotheses become possible. If for example $\mu_{ab} = \mu_a + \mu_b, \mu_{ac} = \mu_a + \mu_c, \mu_{bc} = \mu_b + \mu_c, \mu_{abc} = \mu_a + \mu_b + \mu_c$, $v_1 = v_6 = v_8 = v_{12}, v_9 = v_4, v_{10} = v_7, v_2 = v_5$, and $v_{11} = v_9$ at all ages, we would say that disease $a$ progresses independently of $b$ and $c$. Formally a progressive Markov Illness Death model with $k$ diseases will be a process in which there is a positive intensity for a transition from state $\alpha_1$ to state $\alpha_2$ if the disease combination $\alpha_2$ contains exactly one more disease than $\alpha_1$. Let $\Gamma$ be the set of such pairs $(\alpha_1, \alpha_2)$. All other instantaneous transitions between live states are prohibited. To each live state $\alpha$ there will also correspond a death intensity $\mu_a(t)$. Let $\Delta$ denote the set of live states. There are thus $2^k$ death intensities and $2^{k-1}$ transitions between live states. The model contains a total of $2^{k-1}(k + 2)$ age-dependent parameters to be estimated.

Definition

Let $\Theta$ be a set of parameters and $P(x|\theta), x \in \Omega, \theta \in \Theta$ be a family of distributions with sample space $\Omega$, then the parameters
are nonidentifiable iff for every $\theta \in \Theta$ there is $\eta \in \Theta$, $\theta \neq \eta$ such that $P(x|\theta) = P(x|\eta) \forall x \in \Omega$. With this definition the following theorem may be stated:

**Theorem 1.** In the progressive Markov Illness Death model with data from survival and serial sacrifice the age-dependent death intensities $\mu_a(t)$, $a \in \Delta$ are identifiable, but for $k > 1$ the age-dependent intensities of transitions between live states $\nu_j(t)$, $\gamma \in \Gamma$ are nonidentifiable.

Before proving the theorem we will consider an example involving only two diseases and therefore eight age-dependent parameters. The network is illustrated in Fig. 2. Four different functional forms are chosen for $\nu_i(t)$:

- **(i)** $\nu_1(t) = 1$
- **(ii)** $\nu_2(t) = 2 - e^{-t}$
- **(iii)** $\nu_3(t) = 2 - e^t$, $t < \log 2$; zero otherwise.
- **(iv)** $\nu_4(t)$ an arbitrary function between (ii) and (iii).

The units are events per animal per 250 days.

For each of these functions $\nu_2$, $\nu_3$, and $\nu_4$ are taken to be

- $\nu_2(t) = 2 - \nu_1(t)$
- $\nu_3(t) = 1 + (\nu_1(t) - 1)/(e^t - 1)$
- $\nu_4(t) = 1 + (\nu_2(t) - 1)/(1 - e^{-t})$

and for death rates we have $\mu_H = \mu_2 = 1$, $\mu_b = 3$, and $\mu_{ab} = 0.5$ in the same units. The $\nu$'s are drawn in Fig. 3. It is claimed that the four sets of parameters i, ii, iii, and iv cannot be distinguished by a survival and serial sacrifice experiment. Let us first of all consider the consequences of this ambiguity. Suppose we wish to estimate the expected duration of life $l_b$ for an animal, given that it has disease $b$; for cases in which $\nu_4$ is constant we have

$$l_b = \frac{1}{\mu_b + \nu_4} + \frac{\nu_4}{\mu_b + \nu_4}.\frac{1}{\mu_{ab}}.$$

When the true value of $\nu_4$ is zero (set ii), we have $l_b = \frac{1}{\mu_4}$ (83 days). In set i with $\nu_4 = 1$, we have $l_b = \frac{\mu_4}{\mu_b + \nu_4}$ (188 days). Thus, there is no way of resolving ambiguities even of this magnitude by survival and serial sacrifice experiments.

**Proof of Theorem 1.** The random variables in a survival and serial sacrifice experiment consist of lifespans for animals that die and indicators of the diseases present for these animals and the animals that are sacrificed. Let $P_a(t)$ be the probability that an animal is alive and in state $a$ at time $t$. With arbitrarily large small samples we may determine $P(t)$, the probability of being alive at time $t$: $P_a(t)/P(t)$, the probability of having disease combination $\alpha$ given animal is alive at $t$ (by sacrificing); and $\int [\mu_a(x) P_a(x)] dx$, the crude death function for disease combination $\alpha$. From these $\mu_a(x)$ and $P_a(x)$ can be determined and are therefore identifiable.

Now the joint distribution of the random variables in a sacrifice experiment has a density proportional to

$$\prod_{\alpha \in \Delta} \prod_{j=1}^{n_\alpha} P_a(s_{\alpha j}) \prod_{j=1}^{m_\alpha} \mu_a(t_{\alpha j}) P_a(t_{\alpha j})$$

where $s_{\alpha j}$ is the time at which animals were found to be in state $\alpha$ and $t_{\alpha j}$, $j = 1 \ldots m_\alpha$, are the death times at which animals were found to be in state $\alpha$. The joint distribution is consequently a function of $\mu$ and $P$ alone. To show that the $\nu$'s are nonidentifiable it is sufficient to show that
the functions $P_x(t)$, $\alpha \in \Delta$, $t > 0$ do not specify $e$, $\gamma \in \Gamma$ uniquely. Now $P_x(t)$, $\alpha \in \Delta$ satisfies a system of linear differential equations involving the $e$'s and $\gamma$'s [given by Chiang (6), for example]. We may rewrite these equations as

$$A x(t) = b(t)$$

where

$$a_{xy} = \begin{cases} 1 & \text{if transition } \gamma \text{ leads into state } \alpha \\ -1 & \text{if transition } \gamma \text{ leads out of state } \alpha \\ 0 & \text{if transition } \gamma \text{ and state } \alpha \text{ are not connected} \end{cases}$$

in which $\alpha(\gamma)$ is the state from which the transition $\gamma$ originates and $A$ is a $2^k$ by $k2^{k-1}$ dimensional matrix whose elements $a_{xy}$ are given by

Since the sum of the rows of $A$ is zero, the rank of $A$ is at most $2^k - 1$. For $k > 1$ this implies that there exists a vector $u$ such that $Au = 0$. It follows that $x(t) + e(t)u$ is a nonidentifiable solution of Eq. 2, positive for $e(t) > 0$, sufficiently small, since the elements of $x(t)$ are presumed positive, q.e.d.

For the case of two diseases the "gap" between what is observable and what is estimable is small. To fill this gap more than the single observation provided by the post mortem examination will be required per animal. For example, let us suppose that the interaction of leukemia and other diseases is to be investigated, then all parameters can be identified if additional data on leukemia incidence are obtained by taking serial blood samples from randomly selected animals. In general intensities are identifiable when they are assumed constant and sacrifices are made on a number of different days as in the experiments of Upton et al. (3). However, when all sacrifices occur on the same day this is not so and the situation is analogous to the nonidentifiability problem noted by Fix and Neyman (5). Their solution to the problem is similar to the above.

It should be noted that even with the demonstrated ambiguity of the survival and serial sacrifice experiment, there are cases in which the hypothesis of disease independence can be rejected when it is false, because independence implies relationships between the $\mu$'s ($\mu_{ab} = \mu_a + \mu_b$, etc.) and the $P$'s, that is between identifiable parameters. However, failure to reject the hypothesis when it is false may be due not to statistical error of type II but rather to the ambiguity introduced by nonidentifiability.

Concluding remarks

The connection between the Markov Illness Death model and compartment models in general has been noted. The matrix $A$ which appears in the proof of Theorem 1 is the incidence matrix of a directed graph depicted for example in Fig. 1. Evidently results of application to the general class of compartment models are possible.

This work was done, with the partial support of the U.S. Energy Research and Development Administration, and the joint support of the National Institutes of Health, Research Grant ES01299-13, while the author was on leave of absence from Oxford University, England.