Projecting Disease When Death Is Likely

Donald R. Hoover,¹ Yun Peng,¹ Alfred J. Saah,¹ Roger R. Detels,² Charles R. Rinaldo, Jr.,³ and John P. Phair⁴

Projecting disease incidence, prevalence, and net morbidity is often needed when individuals are likely to die, either disease free or after the disease has developed. Examples of this include remission of cancer or heart disease in elderly people who can die from these or other causes and occurrence of a particular acquired immune deficiency syndrome illness in human immunodeficiency virus type 1 (HIV-1) disease. Death is not an ancillary event but, rather, indicates either an end to disease morbidity or an end to risk to ever develop that disease. Thus, time to disease survival analyses that censor disease-free individuals at death can produce misleading results. This paper describes several useful quantifications of disease and death for this setting. A paradigm that utilizes Kaplan-Meier functions to estimate these quantities is introduced. The approach anchors on a four-stage disease/death model: stage A, living without disease; stage B, dead without ever developing disease; stage C, developed the disease and living; and stage D, dead after developing the disease. An application is made to projecting cytomegalovirus disease in a cohort of HIV-1-infected users of zidovudine and Pneumocystis prophylaxis from the Multicenter AIDS Cohort Study (MACS) during 1989–1993.

At 3 years after a CD4⁺ count below 100/µl, a man had an 18.7%, 46.3%, 5.3%, or 29.9% chance, respectively, to be in stage A, B, C, or D. This man, on average, had 0.28 years of cytomegalovirus morbidity during these 3 years.

In this setting, estimates of disease occurrence and morbidity are needed to 1) evaluate current and planned clinical practice, 2) allocate resources, 3) develop screening or counseling programs, and 4) plan clinical trials for treatment and prophylaxis. Depending on the specific question of interest, different types of estimates will be needed. For example, consider 1,000 people infected with HIV-1 being monitored for cytomegalovirus disease. If the goal is to allocate resources for the prophylaxis against or diagnosis of the initial cytomegalovirus disease manifestation, then the incidence of this disease (either as a fraction of the original 1,000 or as a fraction of those alive at a given time) is needed. If the goal is to allocate resources for continuing treatment, then the prevalence of the disease (either as a fraction of the original cohort or among those still living at a given time) is needed.

In this setting, estimates of disease occurrence and morbidity are needed to 1) evaluate current and planned clinical practice, 2) allocate resources, 3) develop screening or counseling programs, and 4) plan clinical trials for treatment and prophylaxis. Depending on the specific question of interest, different types of estimates will be needed. For example, consider 1,000 people infected with HIV-1 being monitored for cytomegalovirus disease. If the goal is to allocate resources for the prophylaxis against or diagnosis of the initial cytomegalovirus disease manifestation, then the incidence of this disease (either as a fraction of the original 1,000 or as a fraction of those alive at a given time) is needed. If the goal is to allocate resources for continuing treatment, then the prevalence of the disease (either as a fraction of the original cohort or among those still living at a given time) is needed.

In this situation, death is informative. Death in a person without (cytomegalovirus) disease indicates that the person will never require resources for prophylaxis, diagnosis, or treatment of that disease. Death in a person with disease means that the person has been removed (by that disease or by other causes) from the population that needs to be treated.

Few applications of relevant incidence and prevalence estimates in this setting have been made. Com-
Cometing risk formulations (1–3) can simultaneously model transition into different disease and death states. However, they require strong assumptions of independence (including Markovian transitions) that will be impossible to validate from the data (4). They also need software that is not available to medical researchers and may not be computationally feasible.

This paper defines relevant estimates of disease incidence and prevalence for a population at high risk of death from other causes. A four-stage transition model illustrates these quantities and is a starting point for developing estimates. An application is made to evaluating cytomegalovirus disease occurrence and morbidity in HIV-1-infected men.

MATERIALS AND METHODS

The disease and death model

Disease and death transition is a simple four-stage process shown in figure 1. All subjects start in stage A (living without the disease). With time, they move to either stage B (dead without ever developing the disease) or stage C (living, having developed the disease). From stage C they move to stage D (dead after developing the disease).

Longitudinal quantities of interest

The longitudinal quantities of interest with respect to time, \( t \), can be summarized in terms of the probabilities of being in each of four stages in figure 1: \( P^{0}_{\text{DIS, ALIVE}} \), the probability of being in stage A (alive without disease) at time \( t \); \( P^{0}_{\text{DIS, DEAD}} \), the probability of being in stage B (dead without having developed the disease) at time \( t \); and \( P^{0}_{\text{DIS, ALIVE}} \), the probability of being in stage C (alive having developed the disease) at stage D (dead after developing the disease) at time \( t \). The sum of stages C and D gives \( P^{0}_{\text{DEAD}} \), the probability of having developed the disease at or before time \( t \). The sum of stages B and D gives \( P^{0}_{\text{DEAD}} \), the probability of having died at or before time \( t \). The ratio at time \( t \) of (stage C/(stage C + stage A)) gives \( P^{0}_{\text{DIS, ALIVE}} \), the probability of having disease for a person living at time \( t \).

Each of the previously described quantities can be estimated by reducing the four-stage process to an appropriate two-stage subprocess and applying currently available Kaplan-Meier or binomial software.

Kaplan-Meier estimation of \( P^{0}_{\text{DIS, ALIVE}} \) from a two-stage submodel

For example, to estimate \( P^{0}_{\text{DIS, ALIVE}} \), the nondiseased stages (A and B) could be merged together into one stage and the two diseased stages (C and D) together into another (figure 2). The transition from the merged nondiseased stage to the merged diseased stage could then be estimated with a Kaplan-Meier approach.

For each study subject, \( i \), let \( t_{2,i} \) be the time from the study entry until the end of the study (or when the data are analyzed), and let \( t_{1,i} \) be the time from study entry until onset of disease. If the person never develops the disease (i.e., dies first), then mathematically \( t_{1,i} \) could be considered to equal infinity. Furthermore, \( t_{2,i} \) is defined whether or not the subject was observed to die during the study period. Let \( t_{i} \) be the minimum of \( (t_{1,i}, \text{ and } t_{2,i}) \), and let \( \sigma_{i} = (1 \text{ if disease was observed before}
FIGURE 2. The four-stage disease and death process subsumed into a two-stage nondiseased, diseased process.

The four-stage disease and death process is shown in the diagram. The stages are:

- **Nondiseased Stage**
  - Alive without Disease
  - Dead without Disease

- **Diseased Stage**
  - Alive with Disease
  - Dead following Disease

The end of study, i.e., \( t_{1,i} < t_{2,i} \) and 0 otherwise. This is analogous to 0 = censored and 1 = not censored. If no dropout (loss to follow-up) occurs before the end of study date, the standard Kaplan-Meier estimate applied to these \((t_{i}, \sigma_{i})\) pairs obtains an appropriate estimate \( \hat{P}^{(0)}_{\text{dis}} \) for \( P^{(0)}_{\text{dis}} \).

To demonstrate this, consider the following simple example of two groups taken from two cities, each with 100 HIV-1-infected men with follow-up beginning January 1, 1990. Assume that, for all men, complete data from registries are available on the onset of cytomegalovirus disease through December 31, 1991 (i.e., \( t_{2,i} = 2 \) years). In each city, 15 men developed cytomegalovirus disease during 1990 (a total of 30 men), and 15 men did so in 1991 (a total of 30 men).

Clearly, 15 percent (30 men) of the original 200 men developed the disease in the first year after study entry, 1990, and 15 percent (30 men) did so in the second year, 1991. This can also be derived from a Kaplan-Meier estimator, as is shown below in an equivalent life table format subdivided into the 0- to 1-year interval and into the 1- to 2-year interval. The men from each city are segregated in the numerators and denominators, as the two cities will be treated differently in subsequent modifications to the example:

1. \[
\frac{100 - 15}{100} + \frac{100 - 15}{100} = 0.85
\]
do not develop disease in year 1;

2. \[
0.85 \times \frac{100 - 15}{100 - 15} + \frac{100 - 15}{100 - 15} = 0.70
\]
do not develop disease in the first 2 years.

The only removal from the denominator in modification 2 was the 30 men (15 from each city) who had developed cytomegalovirus disease in the first year. However, the Kaplan-Meier estimator remains accurate when censoring times vary. To see this, assume that registry information in one city exists only from January 1 to December 31, 1990. Then, \( t_{2,i} \) is 1 year for the 100 men from that city and 2 years for the 100 men from the other city. We observe the 15 cytomegalovirus disease cases from the first year in the former city, but we have no follow-up in that city for the second year. The earlier censoring of 85 men who did not develop cytomegalovirus disease in the first year from one of the identical cities (reflected by their removal from the numerator and denominator of modification 2) does not change the Kaplan-Meier estimate.

2. \[
0.85 \times \frac{100 - 15}{100 - 15} + \frac{100 - 15}{100 - 15} = 0.70
\]
do not develop disease in the first 2 years.

Provided no dropout occurs before the end of the study, basing \( t_{2,i} \) on the planned (designed) follow-up time unbiasedly reflects the time for observing the event. Intuitively, it might seem as if this Kaplan-Meier estimator could be improved by incorporation of knowledge of death without disease. However, to extend \( t_{2,i} \) from knowledge of death without disease (i.e., \( t_{1,i} = \infty \)) would make \( t_{2,i} \) dependent on \( t_{1,i} \) (i.e., longer when \( t_{1,i} = \infty \)) and bias the ensuing Kaplan-Meier estimate. For example, suppose that, in the city censored at 1 year, 70 of the 85 men who had not developed cytomegalovirus disease in the first year had already died during that first year. One would be tempted to let \( t_{2,i} = \infty \) for those 70 men who will never develop disease, leaving only 15 men from that city censored at 1 year. The results would be to overestimate that

\[
0.85 \times \frac{100 - 15}{100 - 15} + \frac{100 - 15}{100 - 15} = 0.768
\]
do not develop disease in the first 2 years. There are no other means to incorporate knowledge of death without disease into the Kaplan-Meier estimator without biasing the results.

However, it should be noted that, if individuals in stage A can be lost to follow-up before the end of the study, then the \( t_{2,i} \) we propose will not reflect their actual censor time. In addition, a \( t_{2,i} \) cannot be derived for men remaining in the combined nondiseased stage.
without creating a biased relation between \( t_{2,i} \) and \( t_{1,i} \), which will invalidate this Kaplan-Meier estimate.

**Kaplan-Meier estimation for \( P_{\text{DIS+DEAD}}^{(0)} \), \( P_{\text{DIS+DEAD}}^{(0)} \) and \( P_{\text{DEAD}}^{(0)} \)**

\( P_{\text{DIS+DEAD}}^{(0)} \): Death without disease is stage B. For a subject who enters stage B, \( t_i \) is \( t_{1,i} \), the time stage B was entered, and \( \sigma_i = 1 \). For a subject whose last stage during the study was stage A, C, or D, \( t_i \) is \( t_{2,i} \), the time from study entry until the end of the study, and \( \sigma_i = 0 \). (However, for this \( t_{2,i} \) to be valid, persons in stages A and C must not drop out before the end of study.) The estimate \( \hat{P}_{\text{DIS+DEAD}}^{(0)} \) obtains by subtracting from 1 the Kaplan-Meier survival estimate on these \((t_i, \sigma_i)\) pairs.

\( P_{\text{DEAD}}^{(0)} \): Death is stages B and D. For those who enter these stages, \( t_i \) is \( t_{1,i} \), the time of death, and \( \sigma_i = 1 \). If follow-up is complete to the end of the study, then \( t_i = t_{2,i} \), the time from study entry until the end of the study, and \( \sigma_i = 0 \). (However, for this \( t_{2,i} \) to be valid, individuals in stages A and C must not drop out before the end of study.) Subtracting from 1 the Kaplan-Meier survival estimate on these \((t_i, \sigma_i)\) pairs gives \( \hat{P}_{\text{DIS+DEAD}}^{(0)} \).

For a fixed time \( t \), the estimates \( \hat{P}_{\text{DEAD}}^{(0)} \) and \( \hat{P}_{\text{DIS+ALIVE}}^{(0)} \) are independent. \( \hat{P}_{\text{DEAD}}^{(0)} \) is determined only by the number and timing of deaths and censorings at and prior to \( t \). \( \hat{P}_{\text{DIS+ALIVE}}^{(0)} \) is based only on subjects not dead or censored at or prior to \( t \). Because of the independence of individuals, the health status of a person alive and under follow-up at time \( t \) is independent of the number of deaths and censorings prior to \( t \). Since the two estimators on the right-hand side of formula 1 are independent, by Greenwood’s (5) formula:

\[
\text{var}(\hat{P}_{\text{DIS+ALIVE}}^{(0)}) \approx (1 - \hat{P}_{\text{DEAD}}^{(0)})^2 \text{var}(\hat{P}_{\text{DIS+ALIVE}}^{(0)}) + (\hat{P}_{\text{DIS+ALIVE}}^{(0)})^2 \text{var}(\hat{P}_{\text{DEAD}}^{(0)})
\]

If the disease is not curable, then the expected amount of disease morbidity time is the integral of the proportion diseased and alive:

\[
\text{[Expected time with disease morbidity]} \equiv \int \hat{P}_{\text{DIS+ALIVE}}^{(0)} dt. \tag{2}
\]

**A commonly misused estimator of disease incidence**

A Kaplan-Meier estimate for the cumulative incidence of disease that censors disease-free individuals at death is often used. We denote this estimate as \( P_{\text{DIS+CN+DTH}}^{(0)} \). For example, let 40 men be studied 2 years for the occurrence of cytomegalovirus disease. Suppose that 30 men die without the disease in the first year and that 10 others develop the disease during the second year. Then, \( P_{\text{DIS+CN+DTH}}^{(0)} \) will be 1.0. The 30 first year deaths are censored, so only 10 men remained “under follow-up” in the second year, all of whom developed disease. In contrast to this estimate, only 0.25 (10/40) of the original cohort actually developed the disease. In the Discussion, we describe how \( P_{\text{DIS+CN+DTH}}^{(0)} \) has been interpreted in the literature and why these interpretations may be misleading, and we show how the estimates derived here may be more appropriate.

**RESULTS**

Monitoring HIV-1-infected persons with CD4\(^+\) cell counts below 100/\(\mu\)l for cytomegalovirus disease is now recommended (6, 7). Issues exist about cost effectiveness of prophylactic and therapeutic regimens against this disease (8–11). Interactions of these medications with other anti-acquired immune deficiency syndrome (AIDS) drugs and cumulative toxicities are also of concern. Estimates on the timing of death and
cytomegalovirus disease may help to answer questions about the cost effectiveness of monitoring, treating, and preventing this disease.

We evaluated 351 HIV-1-seropositive gay men from the Multicenter AIDS Cohort Study (MACS) (12) who used zidovudine and Pneumocystis carinii prophylaxis and had CD4+ cell counts fall below 100/µl. Dates of cytomegalovirus disease and death were determined with active and passive surveillance. Participants had CD4+ cell counts measured every 6 months. Acquisition of these events is essentially complete even among men who quit study participation (13). However, the men rarely discontinued study participation once immunosuppression became this poor (i.e., CD4 counts below 100/µl). Other analyses from this cohort also censor (window) subjects at the date of analysis (14, 15). Follow-up here begins at the date the CD4+ count was first measured below 100/µl (CD4100) and ends on December 31, 1993.

Figure 3 gives the probabilities over time to be in each cytomegalovirus disease and death stage shown in figure 1. At time \( t = 0 \), all participants start in the middle unshaded region of figure 3, alive without cytomegalovirus disease (stage A in figure 1). From here, men can move to stage B, dead without diagnosis of cytomegalovirus disease. This is the upper shaded region of figure 3. Its upper boundary line is \( P_{DIS}^{0} \), and its lower boundary is \( P_{DIS,DEAD}^{0} \). Transition into stage C, alive after the diagnosis of cytomegalovirus disease, can also occur from stage A. Stage C is the second to lowest shaded region of figure 3. Its upper boundary line is \( P_{DIS,ALIVE}^{0} \), and its lower boundary is \( P_{DIS,DEAD}^{0} \).

Table 1 summarizes the regions in figure 3 at various times. At 1 year, 2 years, and 3 years, respectively, after CD4100, only 68.4 percent, 31.0 percent, and 18.7 percent, respectively, remained alive without cytomegalovirus disease. By these times, 13.0 percent, 30.4 percent, and 35.2 percent, respectively, had developed cytomegalovirus disease. By 1, 2, and 3 years, respectively, after CD4100, 18.6 percent, 38.6 percent, and 46.3 percent, respectively, had died without developing cytomegalovirus disease, while 4.5 percent, 15.8 percent, and 29.9 percent, respectively, had died after developing this disease.

Also shown in figure 3 is a dashed line giving \( P_{DIS,ALIVE}^{0} \), the Kaplan-Meier estimator of cytomegalovirus disease “cumulative incidence” which censors disease-free individuals who die at their date of death. These “cumulative incidences” are 14.1 percent for 1 year, 42.3 percent for 2 years, and 53.5 percent for 3 years. As we describe later, while this type of estimate is often used, we believe it does not have a valid interpretation.

Figure 4a gives the fractions of the original cohort who were living with cytomegalovirus disease at given times, \( P_{DIS,ALIVE}^{0} \), and the fractions initially diag-

\[ \text{FIGURE 3. Longitudinal estimates of transition to cytomegalovirus disease and death.} \]

\[ \square \), stage A (alive without cytomegalovirus disease); \[ \Box \), stage B (dead having never been diagnosed with cytomegalovirus disease); \[ \Diamond \), stage C (alive and diagnosed with cytomegalovirus disease); \[ \blacklozenge \), stage D (dead after diagnosis of cytomegalovirus disease); ---, Kaplan-Meier "cumulative incidence" of cytomegalovirus disease obtained by censoring at death.\]
TABLE 1. Longitudinal estimates of transition to cytomegalovirus disease and death stage with confidence intervals among men receiving zidovudine and Pneumocystis prophylaxis in the Multicenter AIDS* Cohort Study, 1989–1993

<table>
<thead>
<tr>
<th>Stage</th>
<th>Time after CD4&lt;sup&gt;+&lt;/sup&gt; cell counts first fall below 100/μl</th>
<th>1 year</th>
<th>2 years</th>
<th>3 years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (%)</td>
<td>95% CI</td>
<td>Estimate (%)</td>
<td>95% CI</td>
</tr>
<tr>
<td>A (alive without diagnosed cytomegalovirus disease)</td>
<td>68.4</td>
<td>62.6–73.4</td>
<td>31.0</td>
<td>25.3–37.5</td>
</tr>
<tr>
<td>B (dead without diagnosed cytomegalovirus disease)</td>
<td>18.6</td>
<td>14.8–24.0</td>
<td>38.6</td>
<td>32.7–45.1</td>
</tr>
<tr>
<td>C and D (diagnosed with cytomegalovirus disease)</td>
<td>13.0</td>
<td>9.6–17.3</td>
<td>30.4</td>
<td>25.0–36.7</td>
</tr>
<tr>
<td>D (dead after diagnosis of cytomegalovirus disease)</td>
<td>4.5</td>
<td>2.3–7.2</td>
<td>15.8</td>
<td>11.6–21.3</td>
</tr>
</tbody>
</table>

* AIDS, acquired immune deficiency syndrome; CI, confidence interval.

nosed with cytomegalovirus disease within 6 months after that timepoint, \( P_{\text{DIS}}^{(t)} - P_{\text{DIS}}^{(t+6 \text{ months})} \).

At \( t = 0.5, 1.0, 1.5, 2.0, 2.5, \) and 3.0 years, respectively, after CD4<sub>100</sub>, 5.2 percent, 8.5 percent, 12.6 percent, 14.6 percent, 12.0 percent, and 5.6 percent, respectively, of the initial subjects were living with cytomegalovirus disease. Based on these numbers and the trapezoidal rule for integrating formula (2), every 100 persons whose CD4<sup>+</sup> counts fell below 100/μl generated 27.85 person-years, \( 0.052 + 0.085 + 0.126 + 0.146 + 0.120 + \frac{1}{4} (0.0 + 0.056) \), of cytomegalovirus disease during the next 3 years.

Figure 4b shows the fraction of the initial cohort who have an initial manifestation of cytomegalovirus disease during each 6-month interval. Between 0–0.50, 0.51–1.00, 1.01–1.50, 1.51–2.00, and 2.01–2.50 years, respectively, after CD4<sub>100</sub>, 5.3 percent, 7.5 percent, 9.8 percent, 7.1 percent, and 4.3 percent, respectively, of the original cohort developed cytomegalovirus disease.

Figure 5a gives prevalences of cytomegalovirus disease among the men alive at \( t, P_{\text{DIS}}^{(t)}_{\text{ALIVE}} \) (i.e., the ratio of stage C in figure 1 to the sum of stages A and C). Among those men alive at 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 years, respectively, after CD4<sub>100</sub>, 5.3 percent, 7.5 percent, 9.8 percent, 7.1 percent, and 4.3 percent, respectively, of the original cohort developed cytomegalovirus disease.

Figure 5b gives the 6-month future incidences of cytomegalovirus disease among men alive and disease free at 0, 0.5, 1.0, 1.5, and 2.0 years. The denominators of these incidences are not the entire cohort but, rather, the portions of the cohort in stage A at the beginning of the interval. From 0–0.50, 0.51–1.00, and 1.01–1.50 years, respectively, after CD4<sub>100</sub>, the 6-month incidence of cytomegalovirus disease among surviving subjects disease free rose from 5.3 percent to 8.5 percent and then to 14.2 percent. This 6-month incidence remained stable at 15.0 percent and 14.6 percent, respectively, from 1.51–2.00 and 2.01–2.50 years, respectively, after CD4<sub>100</sub>.

**DISCUSSION**

Potential uses of these estimates

Different questions call for different answers. The standard analysis of time to event that censors at death may not be answering a question that is relevant to a given concern. We now discuss several relevant questions about disease incidence and prevalence in a population at risk for death from other causes that one might answer and indicate the appropriate estimates to answer each.

The future timing for initial manifestations of disease is needed to plan screening programs or entry into treatment studies. This information is provided by examining the transition out of stage A into stage C, as shown in figure 4b. For example, about 10 percent of HIV-1-infected homosexual men reaching CD4<sub>100</sub> today will need to begin treatment for an initial cytomegalovirus condition between 1.0 and 1.5 years from now.

Long-term future costs for the care of disease are reflected by \( P_{\text{DIS}}^{(t)}_{\text{ALIVE}} \), the “percentages of the original cohort in stage C, living with disease” at various times, as shown in figure 4a. Of particular economic interest is that, for every 100 HIV-1-infected homosexual men with CD4<sup>+</sup> cell counts falling below...
100/μl today, 27.85 person-years of life with cytomegalovirus disease morbidity can be expected over the next 3 years.

Projecting short-term needs for counseling and treating surviving immunosuppressed patients with initial manifestations of disease is important. This is guided by the incidences of disease among the living in figure 5b. If a person continues to survive disease free more than 1 year beyond the CD4100 threshold, his chances of developing cytomegalovirus disease during the next 6 months remain between 14 and 15 percent.

Prophylaxis against diseases can be both expensive and toxic (11). The effectiveness of a prophylaxis must thus be based on whether it delays disease (stage A), and if it 1) delays (or accelerates) the date of death (stages B and D of figure 1); 2) delays the earliest of the dates of disease and death (stages B and C of figure 1); and 3) delays the date of death attributable to the disease (a subset of stage D in figure 1). The analyses here provide normative data on the rates of these events in subjects not receiving prophylaxis.

A prophylaxis trial starting at CD4100 and lasting 1 year, aimed solely at detecting a protective effect against death due to cytomegalovirus disease, should consider that 13.0 percent of subjects not receiving the experimental agent would be expected to develop the disease and that an additional 18.6 percent of subjects die in that year. However, only 4.1 percent die in the first year after previously developing cytomegalovirus disease. Thus, unless the sample size was huge, that study would have almost no power to detect a prophylaxis effect on death from cytomegalovirus disease.

**Comparisons with a commonly misused estimator of disease incidence**

Many articles that analyze disease onset in settings where death from other causes is likely fit Kaplan-
Meier estimates (or actuarial, Poisson, and Cox model analogs) that censor disease-free people at death (16–24). This estimator is $\hat{\pi}_{\text{DIS-CN-DTH}}(t)$ described in Materials and Methods. It has no direct medical or biologic meaning, but it has been misinterpreted in ways that overestimate the occurrence of disease.

For example, $\hat{\pi}_{\text{DIS-CN-DTH}}(3 \text{ years})$ from the analysis here (figure 3, dashed line) was 53.5 percent. This was almost 20 percent greater than $\hat{\pi}_{\text{DIS}}(3 \text{ years})$, 35.2 percent, the correct estimate for the fraction of the original cohort who had developed cytomegalovirus disease by 3.0 years (from table 1).

Censoring disease-free people as "lost to follow-up" at death mathematically assumes that these dead people (in stage B of figure 1) continue to develop disease (i.e., move into stage C) at the same rate as do the living people in stage A. Dead people (in stage B) can never develop disease (move into stage C or stage D), whereas those alive but lost to follow-up can develop disease. Thus, censoring disease-free people at death overestimates the cumulative probability of disease for the entire cohort.

If the processes leading to disease and death without disease were independent, then $\hat{\pi}_{\text{DIS-CN-DTH}}(t)$ would estimate the proportion who would develop disease by time $t$ if death from other causes could be eliminated. It is, however, generally impossible to determine from a given set of data whether two competing processes are independent (4). Furthermore, independence of these processes is medically unlikely. For example, in

---

**FIGURE 5.** a, prevalences of cytomegalovirus disease among those still living at selected time points (with 95% confidence intervals); b, 6-month incidences of cytomegalovirus disease among those still living and disease free at the beginning of the interval (with 95% confidence intervals), among human immunodeficiency virus type 1-infected users of zidovudine and *Pneumocystis* prophylaxis in the Multicenter AIDS Cohort Study, 1989–1993.
general, sicker individuals may be both more likely to develop the disease and to die of other causes (positive dependence).

Estimates from Kaplan-Meier models that censor at death (i.e., $\hat{P}^{\text{DIS-CN-DTH}}_{\text{DIS-CN-DTH}}(3\text{ years})$) are also often reported as survivor prevalences, the probabilities of those living at a given time, $t$, to have the disease (22–24). Again, $\hat{P}^{3\text{ years}}_{\text{DIS-CN-DTH}}$ from the Kaplan-Meier model that censored at death in the example was 53.5 percent, more than double $\hat{P}^{3\text{ years}}_{\text{DIS-ALIVE}}$. 23.8 percent, the observed prevalence of disease among survivors 3 years after CD4$_{100}$ (figure 5a).

The Kaplan-Meier estimate of the time to disease that censors disease-free individuals at death increments the estimator as transitions from stage A to stage C of figure 1 occur. However, it does not consider transitions to death from stages A and C. Usually the disease is lethal; diseased cases in stage C die more rapidly than those without disease in stage A. The censor disease-free-at-death approach fails to account for this shorter survival in those with disease and thus overestimates disease prevalence among survivors.

Limitations and implications

We have given simple and important quantifications of disease occurrence and morbidity in settings where death from other causes is likely. We have presented an approach that uses available statistical software to estimate these quantities. One data requirement is that participants not drop out before the end of the study. In the example used for illustration here, that was the case. We believe that this is also likely to be true in general, as people at high risk for death are often nonmobile and under intense observation.

Further research toward implementing these concepts with other parametric and nonparametric models may be productive. For example, nonparametric formulations exist to simultaneously estimate the transition from one baseline stage to two (or more) adjacent alternate stages (i.e., multiple decrements) (3). Unlike the Kaplan-Meier model, software to implement these models is not available nor have computationally feasible variances been demonstrated. However, multiple decrement estimation could be incorporated into the paradigm presented here and would eliminate the need for an assumption of no dropout before the end of the study.

Medical researchers have erroneously censored disease-free people at death, consequently overestimating the magnitude of the disease. The concepts developed here and the ability to produce estimates with accessible software may reduce this misapplication and expand the information retrievable on disease occurrence and morbidity.

ACKNOWLEDGMENTS

This work was supported by NIH/NIAID cooperative agreement U01-AI-35043.

The authors acknowledge Dr. Lewis Schrager and Dr. Harout Armenian, as well as Edith Muth for preparation of the manuscript.


REFERENCES

11. Gallant JE, Moore RO, Chaisson RE. Prophylaxis for oppor-