Survival analysis: A primer March, 2008
David A. Freedman UC Berkeley

In this paper, I will discuss life tables and Kaplan-Meier estimators, which are similar to life tables. Then I turn to proportional-hazards models, aka "Cox models." Along the way, I will look at the efficacy of screening for lung cancer, the impact of negative religious feelings on survival, and the efficacy of hormone replacement therapy.

What are the conclusions about statistical practice? Proportional-hazards models are frequently used to analyze data from randomized controlled trials. This is a mistake. Randomization does not justify the models, which are rarely informative. Simpler analytic methods should be used first.

With observational studies, the models would help us disentangle causal relations *if* the assumptions behind the models could be justified. Justifying those assumptions, however, is fraught with difficulty.

Cross-sectional life tables

Cross-sectional life tables date back to John Graunt and Edmond Halley in the 17th century. There were further developments by Daniel Bernoulli in 1760, when he computed what life expectancy would be—if smallpox were eliminated. His calculations make a key assumption, to be discussed later: the independence of competing risks.

Here is a simple discrete case to illustrate the idea behind cross-sectional life tables ("cross-sectional" because they can be computed from vital statistics available at one point in time, covering people of all ages). There are N_t people alive at the beginning of age t, but n_t of them die before reaching age $t+1$. The death probability in year t of life is n_t/N_t, the survival probability is $1 - n_t/N_t$. The probability at birth ("age 0") of surviving T years or more is estimated as

$$\prod_{t=0}^{T-1} \left(1 - \frac{n_t}{N_t}\right). \tag{1}$$

There are corrections to make if you want to get from discrete time to continuous time; this used to be a major topic in applied mathematics. However, the big assumption in constructing the life table is that death rates do not change over time. If there is a trend, the life table will be biased. From Bernoulli's day onwards, death rates have been going down in the Western world; this was the beginning of the demographic transition (Kirk, 1996). Therefore, cross-sectional life tables understate life expectancy.

Hazard rates

Let τ be a positive random variable—the waiting time for failure. Suppose τ has a continuous positive density f. The distribution function is $F(t) = \int_0^t f(u)\,du$, with $F' = f$. The survival function is $S = 1 - F$. The hazard rate is

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{2}$$

The intuition behind the formula is that $h(t)\,dt$ represents the conditional probability of failing in the interval $(t, t + dt)$, given survival until time t.

We can recover f, S, and F from the hazard rate:

$$S(t) = 1 - F(t) = \exp\Big(-\int_0^t h(u)\,du\Big), \tag{3}$$

$$f(t) = h(t)S(t). \tag{4}$$

It follows from (2) or (3) that $\int_0^\infty h(u)\,du = \infty$. In many studies, however, the failure rate is low. Then $F(t) \approx 0$, $S(t) \approx 1$, and $f(t) \approx h(t)$ over the observable range of t's.

Technical notes. (i) To derive $\int_0^\infty h(u)\,du = \infty$ from (2): if $0 \le t_n < t_{n+1}$, then

$$\int_{t_n}^{t_{n+1}} h(u)\,du > [S(t_n) - S(t_{n+1})]/S(t_n).$$

Choose the t_n inductively, with $t_0 = 0$ and t_{n+1} so large that $S(t_{n+1}) < S(t_n)/2$. Then sum over n. Also see Rudin (1976, p. 79). The derivation from (3) is clear, again because $S(\infty) = 0$.

(ii) Equation (2) says that $S'/S = -h$. Solving for S with the constraint $S(0) = 1$ gives $S(t) = \exp\big(-\int_0^t h(u)\,du\big)$.

Here are four types of failure, the first two drawn from consulting projects, the others to be discussed later on. (i) A light bulb burns out. (This may seem too trite to be true, but the client was buying a lot of bulbs: which brand to buy, and when to relamp?) (ii) A financial institution goes out of business. (iii) A subject in a clinical trial dies. (iv) A subject in a clinical trial dies of a pre-specified cause, for instance, lung cancer.

Some examples may help to clarify the mathematics.

Example 1. If τ is standard exponential, $P(\tau > t) = \exp(-t)$ is the survival function, and the hazard rate is $h \equiv 1$.

Example 2. If τ is Weibull, the survival function is by definition

$$P(\tau > t) = \exp(-at^b). \tag{5}$$

The density is

$$f(t) = abt^{b-1}\exp(-at^b), \tag{6}$$

and the hazard rate is

$$h(t) = abt^{b-1}. \tag{7}$$

Here, $a > 0$ and $b > 0$ are parameters. The parameter b controls the shape of the distribution, and a controls the scale. If $b > 1$, the hazard rate keeps going up: the longer you live, the shorter your future life will be. If $b < 1$, the hazard rate goes down: the longer you live, the longer your future life will be. The case $b = 1$ is the exponential: if you made it to time t, you still have the same exponential amount of lifetime left ahead of you.

Example 3. If c and d are positive constants and U is uniform on the unit interval, then $c(-\log U)^d$ is Weibull: $a = (1/c)^{1/d}$ and $b = 1/d$.

Example 4. If τ_i are independent with hazard rates h_i, the minimum of the τ's has hazard rate $\sum_i h_i$.

Turn now to the independence of competing risks. We may have two kinds of failure, like death from heart disease or death from cancer. Independence of competing risks means that the time to death from heart disease is independent of the time to death from cancer.

There may be a censoring time $\mathfrak{c}$ as well as the failure time τ. Independence of competing risks means that $\mathfrak{c}$ and τ are independent. The chance that $\tau > t + s$ given $\tau > t$ and $\mathfrak{c} = t$ equals the chance that $\tau > t + s$ given $\tau > t$, without the $\mathfrak{c}$. If they lose track of you, that doesn't change the probability distribution of your time to failure. (Independence of $\mathfrak{c}$ and τ is often presented as a separate condition, rather than being folded into the independence of competing risks.)

The Kaplan-Meier estimator

In a clinical trial, t is usually time on test, that is, time from randomization. Time on test is to be distinguished from age and calendar time ("period"). The analysis here assumes stationarity: failure times are determined by time on test, and are not influenced by age or period.

We also have to consider censoring, which occurs for a variety of reasons. For instance, one subject may withdraw from the study. Another subject may get killed by an irrelevant cause: if failure is defined as death from heart disease, and the subject gets run over by a bus, this is not failure, this is censoring. (At least, that's the party line.) A third subject may be censored because he survived until the end of the study.

Subjects may be censored at late times if they were early entrants to the trial. Conversely, early censoring is probably common among late entrants. We're going to lump all forms of censoring together, and we're going to assume independence of competing risks.

Suppose there are no ties (no two subjects fail at the same time). At any particular time t with a failure, let N_t be the number of subjects on test "at time $t-$," that is, just before time t. The probability of surviving from $t-$ to $t+$ is $1 - 1/N_t$. You just multiply these survival probabilities to get a monotone decreasing function, which is flat between failures but goes down a little bit at each failure:

$$T \to \prod_{t \le T} \Big(1 - \frac{1}{N_t}\Big). \tag{8}$$

This is the Kaplan-Meier (1958) survival curve. Notice that N_t may go down between failures, at times when subjects are censored. However, the Kaplan-Meier curve does not change at censoring times. Of course, censored subjects are excluded from future N_t's, and do not count as failures either. The modification for handling ties is pretty obvious.

In a clinical trial, we would draw one curve for the treatment group and one for the control group. If treatment postpones time to failure, the survival curve for the treatment group will fall off more slowly. If treatment has no effect, the two curves will be statistically indistinguishable.

What is the curve estimating? If subjects in treatment are independent with a common survival function, that is what we will be getting, and likewise for the controls. What if subjects aren't IID? Under suitable regularity conditions, with independent subjects, independence of competing risks, and stationarity, the Kaplan-Meier curve for the treatment group estimates the average curve we would see if all subjects were assigned to treatment. Similarly for the controls.

Kaplan-Meier estimators are subject to bias in finite samples. Technical details behind consistency results are not simple; references will be discussed below. Among other things, the times t at which failures occur are random. The issue is often finessed (in this paper too).

The Kaplan-Meier curve is like a cross-sectional life table, but there is some difference in perspective. The context for the life table is grouped cross-sectional data. The context for the Kaplan-Meier curve is longitudinal data on individual subjects.

How would we estimate the effect on life expectancy of eliminating smallpox? In Bernoulli's place, we might compute the Kaplan-Meier curve, censoring the deaths from smallpox. What he did was to set up differential equations describing the hazard rate ("force of mortality") due to various causes. Independence of competing risks is assumed. If the people who died of smallpox were likely to die shortly thereafter of something else anyway ("frailty"), we would all be over-estimating the impact of eliminating smallpox.

Using data from Halley (1693), Bernoulli estimated that life expectancy at birth was around 27 years; eliminating smallpox would add 3 years to this figure. In 2007, life expectancy at birth was 80 years or thereabouts, in the United States, the United Kingdom, France, Germany, the Netherlands, and many other European countries—compared to 35 years or so in Swaziland and some other very poor countries.

An application of the Kaplan-Meier estimator

If cancer can be detected early enough, before it has metastasized, there may be improved prospects for effective therapy. That is the situation for breast cancer and cervical cancer, among other examples. Claudia Henschke et al (2006) tried to make the case for lung cancer. This was an intriguing but unsuccessful application of survival analysis.

Henschke and her colleagues screened 31,567 asymptomatic persons at risk for lung cancer using low-dose CT (computerized tomography), resulting in a diagnosis of lung cancer in 484 participants. These 484 subjects had an estimated ten-year survival rate of 80%. Of the 484 subjects, 302 had stage I cancer and were resected within one month of diagnosis. The resected group had an estimated ten-year survival rate of 92%. The difference between 92% and 80% was reported as highly significant.

Medical terminology. Cancer has metastasized when it has spread to other organs. Stage describes the extent to which a cancer has progressed. Stage I cancer is early-stage cancer, which usually means small size, limited invasiveness, and a good prognosis. In a resection, the surgeon opens the chest cavity, and removes the diseased portion of the lung. Adenocarcinomas (referred to below) are cancers that appear to have originated in glandular tissue.

Survival curves were computed by the Kaplan-Meier method: see Figure 2 in the paper. Tick marks usually show censoring. Deaths from causes other than lung cancer were censored, but a lot of the censoring is probably because the subjects survived until the end of the study. In this respect among others, crucial details are omitted. The authors conclude that

> "CT screening . . . can detect clinical stage I lung cancer in a high proportion of persons when it is curable by surgery. In a population at risk for lung cancer, such screening could prevent some 80% of deaths from lung cancer." [p. 1769]

The evidence is weak. For one thing, conventional asymptotic confidence intervals on the Kaplan-Meier curve are shaky, given the limited number of data after month 60. (Remember, late entrants to the trial will only be at risk for short periods of time.) For another thing, why are the authors looking only at deaths from lung cancer rather than total mortality? Next, stage I cancers—the kind detected by the CT scan—are small. This augurs well for long-term survival, treatment or no treatment. Even more to the point, the cancers found by screening are likely to be slow-growing. That is "length bias."

Table 3 in Henschke et al shows that most of the cancers were adenocarcinomas; these generally have a favorable prognosis. Moreover, the cancer patients who underwent resection were probably healthier to start with than the ones who didn't. In short, the comparison between the resection group and all lung cancers is uninformative. One of the things lacking in this study is a reasonable control group.

If screening speeds up detection, that will increase the time from detection to death—even if treatment is ineffective. The increase is called "lead time" or "lead-time bias." (To measure the effectiveness of screening, you might want to know the time from detection to death, net of lead time.) Lead time and length bias are discussed in the context of breast cancer screening by Shapiro et al (1988).

When comparing their results to population data, Henschke et al measure benefits as the increase in time from diagnosis to death. This is misleading, as we have just noted. CT scans speed up detection, but we do not know whether that helps the patients live longer, because we do not know whether early treatment is effective. Henschke et al are assuming what needs to be proved. For additional discussion, see Patz et al (2000) and Welch et al (2007).

Lead time bias and length bias are problems for observational studies of screening programs. Well-run clinical trials avoid such biases, if benefits are measured by comparing death rates among those assigned to screening and those assigned to the control group. This is an example of the intention-to-treat principle (Bradford Hill, 1961, p. 259).

A hypothetical will clarify the idea of lead time. "Crypto-megalo-grandioma" (CMG) is a dreadful disease, which is rapidly fatal after diagnosis. Existing therapies are excruciating and ineffective. No improvements are on the horizon. However, there is a screening technique that can reliably detect the disease 10 years before it becomes clinically manifest. Will screening increase survival time from diagnosis to death? Do you want to be screened for CMG?

The proportional-hazards model in brief

Assume independence of competing risks; subjects are independent of one another; there is a baseline hazard rate $h > 0$, which is the same for all subjects. There is a vector of subject-specific characteristics X_{it}, which is allowed to vary with time. The subscript i indexes subjects and t indexes time. There is a parameter vector β, which is assumed to be the same for all subjects and constant over time. Time can be defined in several ways. Here, it means time on test; but see Thiébaut and Bénichou (2004). The hazard rate for subject i is assumed to be

$$h(t)\exp(X_{it}\beta). \tag{9}$$

No intercept is allowed: the intercept would get absorbed into h. The most interesting entry in X_{it} is usually a dummy for treatment status. This is 1 for subjects in the treatment group, and 0 for subjects in the control group. We pass over all technical regularity conditions in respectful silence.

The likelihood function is not a thing of beauty. To make this clear, we can write down the log likelihood function $L(h, \beta)$, which is a function of the baseline hazard rate h and the parameter vector β. For the moment, we will assume there is no censoring and the X_{it} are constant (not random). Let τ_i be the failure time for subject i. By (3)-(4),

$$L(h, \beta) = \sum_{i=1}^{n} \log f_i(\tau_i | h, \beta), \tag{10a}$$

where

$$f_i(t|h, \beta) = h_i(t|\beta) \exp\Big(-\int_0^t h_i(u|\beta)\, du\Big), \tag{10b}$$

and

$$h_i(t|\beta) = h(t) \exp(X_{it}\beta). \tag{10c}$$

This is a mess, and maximizing over the infinite-dimensional parameter h is a daunting prospect.

Cox (1972) suggested proceeding another way. Suppose there is a failure at time t. Remember, t is time on test, not age or period. Consider the set R_t of subjects who were on test just before time t. These subjects have not failed yet, or been censored; so they are eligible to fail at time t. Suppose it was subject j who failed. Heuristically, the chance of it being subject j rather than anybody else in the risk set is

$$\frac{h(t)\exp(X_{jt}\beta)\, dt}{\sum_{i \in R_t} h(t)\exp(X_{it}\beta)\, dt} = \frac{\exp(X_{jt}\beta)}{\sum_{i \in R_t} \exp(X_{it}\beta)}. \tag{11}$$

Subject j is in numerator and denominator both, and by assumption there are no ties: ties are a technical nuisance. The baseline hazard rate $h(t)$ and the dt cancel! Now we can do business.

Multiply the right side of (11) over all failure times to get a "partial likelihood function." This is a function of β. Take logs and maximize to get $\hat{\beta}$. Compute the Hessian—the second derivative matrix of the log partial likelihood—at $\hat{\beta}$. The negative of the Hessian is the "observed partial information." Invert this matrix to get the estimated variance-covariance matrix for the $\hat{\beta}$'s. Take the square root of the diagonal elements to get asymptotic SEs.

Partial likelihood functions are not real likelihood functions. The harder you think about (11) and the multiplication, the less sense it makes. The chance of what event, exactly? Conditional on what information? Failure times are random, not deterministic; this is ignored by (11). The multiplication is bogus. For example, there is no independence: if Harriet is at risk at time T, she cannot have failed at an earlier time t. Still, there is mathematical theory to show that $\hat{\beta}$ performs like a real MLE, under the regularity conditions that we have passed over; also see Example 5 below.

Proportional-hazards models are often used in observational studies and in clinical trials. The latter fact is a real curiosity. There is no need to adjust for confounding if the trial is randomized. Moreover, in a clinical trial, the proportional-hazards model makes its calculations conditional on assignment. The random elements are the failure times for the subjects. As far as the model is concerned, the randomization is irrelevant. Equally, randomization does not justify the model.

A mathematical diversion

Example 5. Suppose the covariates $X_{it} \equiv X_i$ do not depend on t and are nonstochastic; for instance, covariates are measured at recruitment into the trial and are conditioned out. Suppose there is no censoring. Then the partial likelihood function is the ordinary likelihood function for the ranks of the failure times. Kalbfleisch and Prentice (1973) discuss more general results.

Sketch proof. The argument is not completely straightforward, and all the assumptions will be used. As a matter of notation, subject i has failure time τ_i. The hazard rate of τ_i is $h(t)\exp(X_i\beta)$, the density is $f_i(t)$, and the survival function is $S_i(t)$. Let $c_i = \exp(X_i\beta)$. We start with the case $n = 2$. Let $C = c_1 + c_2$. Use (3)-(4) to see that

$$\begin{aligned} P(\tau_1 < \tau_2) &= \int_0^\infty S_2(t) f_1(t)\, dt \\ &= c_1 \int_0^\infty h(t) S_1(t) S_2(t)\, dt \\ &= c_1 \int_0^\infty h(t) \exp\Big(-C\int_0^t h(u)du\Big)\, dt. \end{aligned} \tag{12}$$

Last but not least,

$$C\int_0^\infty h(t)\exp\Big(-C\int_0^t h(u)du\Big)\, dt = 1 \tag{13}$$

by (4). So

$$P(\tau_1 < \tau_2) = \frac{c_1}{c_1 + c_2}, \tag{14}$$

as required.

Now suppose $n > 2$. The chance that τ_1 is the smallest of the τ's is

$$\frac{c_1}{c_1 + \cdots + c_n},$$

as before: just replace τ_2 by $\min\{\tau_2, \ldots, \tau_n\}$. Given that $\tau_1 = t$ and τ_1 is the smallest of the τ's, the remaining τ's are independent and concentrated on (t, ∞). If we look at the random variables $\tau_i - t$, their conditional distributions will have hazard rates $c_i h(t + \bullet)$, so we can proceed inductively. A rigorous treatment might involve regular conditional distributions (Freedman, 1983, pp. 347ff). This completes the sketch proof.

Another argument, suggested by Russ Lyons, is to change the time scale so the hazard rate is identically 1. Under the conditions of Example 5, the transformation $t \to \int_0^t h(u)\,du$ reduces the general case to the exponential case. Indeed, if H is a continuous, strictly increasing function that maps $[0, \infty)$ onto itself, then $H(\tau_i)$ has survival function $S_i \circ H^{-1}$.

The mathematics does say something about statistical practice. At least in the setting of Example 5, and contrary to general opinion, the model does not use time-to-event data. It uses only the ranks: which subject failed first, which failed second, and so forth. That, indeed, is what enables the fitting procedure to get around problems created by the intractable likelihood function.

An application of the proportional-hazards model

Pargament et al (2001) report on religious struggle as a predictor of mortality among very sick patients. Subjects were 596 mainly Baptist and Methodist patients age 55+, hospitalized for serious illness at the Duke Medical Center and the Durham Veterans' Affairs Medical Center. There was two-year followup, with 176 deaths and 152 subjects lost to followup. Key variables of interest were positive and negative religious feelings. There was adjustment by proportional hazards for age, race, gender, severity of illness, . . . , and for missing data.

The main finding reported by Pargament et al is that negative religious feelings increase the death rate. The authors say:

> "Physicians are now being asked to take a spiritual history . . . . Our findings suggest that patients who indicate religious struggle during a spiritual history may be at particularly high risk . . . . Referral of these patients to clergy to help them work through these issues may ultimately improve clinical outcomes; further research is needed . . . ." [p. 1885]

The main evidence is a proportional-hazards model. Variables include age (in years), education (highest grade completed), race, gender, and . . . .

Religious feelings

Positive and negative religious feelings were measured on a seven-item questionnaire, the subject scoring 0–3 points on each item. Two representative items (quoted from the paper) are:

$+$ "collaboration with God in problem solving";
$-$ "decided the devil made this happen."

Physical Health

Number of current medical problems, 1–18.
ADL—Activities of Daily Life. Higher scores mean less ability to function independently.
Patient self-rating, poor to excellent.
Anesthesiologist rating of patient, 0–5 points (0 is healthy, 5 is very sick).

Mental health

MMSE—Mini-Mental State Examination. Higher scores indicate better cognitive functioning.
Depression, measured on a questionnaire with 11 items.
"Quality of life" is observer-rated on five items.

To review briefly, the baseline hazard rate in the model is a function of time t on test; this baseline hazard rate gets multiplied by $e^{X\beta}$, where X can vary with subject and t. Estimation is by partial likelihood.

Table 1 shows estimated hazard ratios, that is, ratios of hazard rates. Age is treated as a continuous variable. The hazard ratio of 1.39 reported in the table is $\exp(\hat{\beta}_A)$, where $\hat{\beta}_A$ is the estimated coefficient for age in the model. The interpretation would be that each additional year of age multiplies the hazard rate by 1.39. This is a huge effect.

Similarly, the 1.06 is $\exp(\hat{\beta}_N)$, where $\hat{\beta}_N$ is the estimated coefficient of the "negative religious feelings" score. The interpretation would be that each additional point on the score multiplies the hazard rate by 1.06.

The proportional-hazards model is linear on the log scale. Effects are taken to be constant across people, and multiplicative rather than additive or synergistic. Thus, in combination, an extra year of age and an extra point on the negative religious feelings scale are estimated to multiply the hazard rate by 1.39×1.06.

The crucial questions. The effect is so small—the hazard ratio of interest is only 1.06—that bias should be a real concern. Was the censoring really independent? Were there omitted variables? Were the measurements too crude? What about reverse causation? For example, there may well be income effects; income is omitted. We might get different answers if age was measured in months rather than years; health at baseline seems to be crudely measured as well. Finally, to illustrate reverse causation, sicker people may have more negative religious feelings.

Table 1. Hazard Ratios. Pargament et al (2001)

Religious feelings −	1.06	$**$
Religious feelings +	0.98	
Age (years)	1.39	$**$
Black	1.21	
Female	0.71	$*$
Hospital	1.14	
Education	0.98	
Physical Health		
Diagnoses	1.04	
ADL	0.98	
Patient	0.71	$***$
Anesthesiologist	1.54	$***$
Mental health		
MMSE	0.96	
Depression	0.95	
Quality of life	1.03	

$*\ P < .10 \quad **\ P < .05 \quad ***\ P < .01$

This is all taken care of by the model. But what is the justification for the model? Here is the authors' answer:

> "This robust semiparametric procedure was chosen for its flexibility in handling censored observations, time-dependent predictors, and late entry into the study." [p. 1883]

The paper has a large sample and a plan for analyzing the data. These positive features are not as common as might be hoped. However—and this is typical—there is scant justification for the statistical model. (The research hypothesis is atypical.)

Does HRT (hormone replacement therapy) prevent heart disease?

There are about 50 observational studies that, on balance, say yes: HRT cuts the risk of heart disease. Several experiments say no: there is no protective effect, and there may even be harm. The most influential of the observational studies is the Nurses' Health Study, which claims a reduction in risk by a factor of 2 or more.

Nurses' Health Study: Observational

Results from the Nurses' Health Study have been reported by the investigators in numerous papers. We consider Grodstein, Stampfer, Manson et al (1996). In that paper, 6224 post-menopausal women on combined HRT are compared to 27,034 never-users. (Former users are considered separately.) There are 0–16 years of followup, with an average of 11 years. Analysis is by proportional hazards. Apparently, failure was defined as either a non-fatal heart attack, or death from coronary heart disease.

The treatment variable is HRT. The investigators report 17 confounders, including age, age at menopause, height, weight, smoking, blood pressure, cholesterol, . . . , exercise. Eleven of the confounders make it into the main model. Details are a little hazy, and there may be some variation from one paper to another. The authors say:

> "Proportional-hazards models were used to calculate relative risks and 95 percent confidence intervals, adjusted for confounding variables. . . . We observed a marked decrease in the risk of major coronary heart disease among women who took estrogen with progestin, as compared with the risk among women who did not use hormones (multivariate adjusted relative risk 0.39; 95 percent confidence interval, 0.19 to 0.78). . . ." [p. 453]

The authors do not believe that the protective effect of HRT can be explained by confounding:

> "Women who take hormones are a self-selected group and usually have healthier lifestyles with fewer risk factors. . . . However, ... participants in the Nurses' Health Study are relatively homogeneous. . . . Unknown confounders may have influenced our results, but to explain the apparent benefit on the basis of confounding variables, one must postulate unknown risk factors that are extremely strong predictors of disease and closely associated with hormone use." [p. 458]

Women's Health Initiative: Experimental

The biggest and most influential experiment is WHI, the Women's Health Initiative. Again, there are numerous papers, but the basic one is Rossouw et al (2002). In the WHI experiment, 16,608 post-menopausal women were randomized to HRT or control. The study was stopped early, with an average followup period of only five years, because HRT led to excess risk of breast cancer.

The principal result of the study can be summarized as follows. The estimated hazard ratio for CHD (Coronary Heart Disease) is 1.29, with a nominal 95% confidence interval of 1.02 to 1.63: "nominal" because the confidence level does not take multiple comparisons into account. The trialists also reported a 95%-confidence interval from 0.85 to 1.97, based on a Bonferroni correction for multiple looks at the data.

The analysis is by proportional hazards, stratified by clinical center, age, prior disease, and assignment to diet. (The effects of a low-fat diet were studied in another, overlapping experiment.) The estimated hazard ratio is $\exp(\hat{\beta}_T)$, where $\hat{\beta}_T$ is the coefficient of the treatment dummy. The confidence intervals are asymmetric because they start on the log scale: the theory produces confidence intervals for β_T, but the parameter of interest is $\exp(\beta_T)$. So you have to exponentiate the endpoints of the intervals.

For a first cut at the data, let us compare the death rates over the followup period (per woman randomized) in the treatment and control groups:

$$\begin{aligned}&231/8506 = 27.2/1000 \text{ vs } 218/8102 = 26.9/1000,\\ &\text{crude rate ratio} = 27.2/26.9 = 1.01.\end{aligned}$$

HRT does not seem to have much of an effect.

The trialists' primary endpoint was CHD. We compute the rates of CHD in the treatment and control groups:

$$\begin{aligned}&164/8506 = 19.3/1000 \text{ vs } 122/8102 = 15.1/1000,\\ &\text{crude rate ratio} = 19.3/15.1 = 1.28.\end{aligned}$$

MI (myocardial infarction) means the destruction of heart muscle due to lack of blood—a heart attack. CHD is coronary heart disease, operationalized here as fatal or non-fatal MI. The rate ratios are "crude" because they are not adjusted for any imbalances between treatment and controls groups.

If you want SEs and CIs for rate ratios, use the delta method, as explained in the appendix. On the log scale, the delta method gives an SE of $\sqrt{1/164 + 1/122} = 0.12$. To get the 95% confidence interval for the hazard ratio, multiply and divide the 1.28 by $\exp(2 \times 0.12) = 1.27$. You get 1.01 to 1.63 instead of 1.02 to 1.63 from the proportional-hazards model. What did the model bring to the party?

Our calculation ignores blocking and time-to-event data. The trialists have ignored something too: the absence of any logical foundation for the model. The experiment was very well done. The data summaries are unusually clear and generous. The discussion of the substantive issues is commendable. The modeling, by contrast, seems ill-considered—although it is by no means unusual. (The trialists did examine the crude rate ratios.)

Agreement between crude rate ratios and hazard ratios from multivariate analysis is commonplace. Indeed, if results were substantively different, there would be something of a puzzle. In a large randomized controlled experiment, adjustments should not make much difference, because the randomization should balance the treatment and control groups with respect to prognostic factors. Of course, if P is close to 5% or 1%, multivariate analysis can push results across the magic line, which has some impact on perceptions.

Were the observational studies right, or the experiments?

If you are not committed to HRT or to observational epidemiology, this may not seem like a difficult question. However, efforts to show the observational studies got it right are discussed in three journals:

International Journal of Epidemiology 2004; 33 (3),
Biometrics 2005; 61 (4),
American Journal of Epidemiology 2005; 162 (5).

For the Nurses' study, the argument is that HRT should start right after menopause, whereas in the WHI experiment, many women in treatment started HRT later. The WHI investigators ran an observational study in parallel with the experiment. This observational study showed the usual benefits. The argument here is that HRT creates an initial period of risk, after which the benefits start. Neither of these timing hypotheses is fully consistent with the data, nor are the two hypotheses entirely consistent with each other (Petitti and Freedman, 2005). Results from late followup of WHI show an increased risk of cancer in the HRT group, which further complicates the timing hypothesis (Heiss et al, 2008).

For reviews skeptical of HRT, see Petitti (1998, 2002). If the observational studies got it wrong, confounding is the likely explanation. An interesting possibility is "prevention bias" or "complier bias" (Barrett-Connor, 1991; Petitti, 1994). In brief, subjects who follow doctors' orders tend to do better, even when the orders are to take a placebo. In the Nurses' study, taking HRT seems to be thoroughly confounded with compliance.

In the clofibrate trial (Freedman-Pisani-Purves, 2007, pp. 14, A-4), compliers had half the death rate of non-compliers—in the drug group and the placebo group both. Interestingly, the difference between compliers and non-compliers could not be predicted using baseline risk factors.

Another example is the HIP trial (Freedman, 2005, pp. 4–5). If you compare women who accepted screening for breast cancer to women who refused, the first group had a 30% lower risk of death from causes other than breast cancer. Here, the compliance effect can be explained, to some degree, in terms of education and income. Of course, the Nurses' Health Study rarely adjusts for such variables.

Many other examples are discussed in Petitti and Chen (2008). For instance, using sunblock reduces the risk of heart attacks by a factor of 2; this estimate is robust when adjustments are made for covariates.

Women who take HRT are women who see a doctor regularly. These women are at substantially lower risk of death from a wide variety of diseases (Grodstein et al, 1997). The list includes diseases where HRT is not considered to be protective. The list also includes diseases like breast cancer, where HRT is known to be harmful. Grodstein et al might object that, in their multivariate proportional-hazards model, the hazard ratio for breast cancer isn't quite significant—either for current users or former users, taken separately.

Simulations

If the proportional-hazards model is right or close to right, it works pretty well. Precise measures of the covariates are not essential. If the model is wrong, there is something of a puzzle: what is being estimated by fitting the model to the data? One possible answer is the crude rate ratio in a very large study population. We begin with an example where the model works, then consider an example in the opposite direction.

The model works

Suppose the baseline distribution of time to failure for untreated subjects is standard exponential. There is a subject-specific random variable W_i which multiplies the baseline time and gives the time to failure for subject i if untreated. The hazard rate for subject i is therefore $1/W_i$ times the baseline hazard rate. By construction, the W_i are independent and uniform on $[0, 1]$. Treatment doubles the failure time, that is, cuts the hazard rate in half—for every subject. We censor at time 0.10, which keeps the failure rates moderately realistic.

We enter $\log W_i$ as the covariate. This is exactly the right covariate. The setup should be duck soup for the model. We can look at simulation data on 5000 subjects, randomized to treatment or control by the toss of a coin. The experiment is repeated 100 times.

> The crude rate ratio is 0.620 ± 0.037. (In other words, the average across the repetitions is 0.620, and the SD is 0.037.)
>
> The model with no covariate estimates the hazard ratio as 0.581 ± 0.039.
>
> The model with the covariate $\log W_i$ estimates the hazard ratio as 0.498 ± 0.032.

The estimated hazard ratio is $\exp(\hat{\beta}_T)$, where $\hat{\beta}_T$ is the coefficient of the treatment dummy in the fitted model. The "real" ratio is 0.50. If that's what you want, the full model looks pretty good. The no-covariate model goes wrong because it fails to adjust for $\log W_i$. This is complicated: $\log W_i$ is nearly balanced between the treatment and control groups, so it is not a confounder. However, without $\log W_i$, the model is no good: subjects do not have a common baseline hazard rate. The Cox model is not "collapsible."

The crude rate ratio (the failure rate in the treatment arm divided by the failure rate in the control arm) is very close to the true value, which is

$$\frac{1 - E[\exp(0.05/W_i)]}{1 - E[\exp(0.10/W_i)]}. \tag{15}$$

The failure rates in treatment and control are about 17% and 28%, big enough so that the crude rate ratio is somewhat different from the hazard ratio: $1/W_i$ has a long, long tail. In this example and many others, the crude rate ratio seems to be a useful summary statistic.

The model is somewhat robust against measurement error. For instance, suppose there is a biased measurement of the covariate: we enter $\sqrt{-\log W_i}$ into the model, rather than $\log W_i$. The estimated hazard ratio is 0.516 ± 0.030, so the bias in the hazard ratio—created by the biased measurement of the covariate—is only 0.016. Of course, if we degrade the measurement further, the model will perform worse. If the covariate is $\sqrt{-\log W_i} + \log U_i$ where U_i is an independent uniform variable, the estimate is noticeably biased: 0.574 ± 0.032.

The model does not work

We modify the previous construction a little. To begin with, we drop W_i. The time to failure if untreated (τ_i) is still standard exponential, and we still censor at time 0.10. As before, the effect of treatment is to double τ_i, which cuts the hazard rate in half. So far, so good: we are still on home ground for the model.

The problem is that we have a new covariate,

$$Z_i = \exp(-\tau_i) + cU_i, \tag{16}$$

where U_i is an independent uniform variable and c is a constant. Notice that $\exp(-\tau_i)$ is itself uniform. The hapless statistician in this fable will have the data on Z_i, but will not know how the data were generated.

The simple proportional-hazards model, without covariates, matches the crude rate ratio. If we enter the covariate into the model, all depends on c. Here are the results for $c = 0$.

The crude rate ratio is 0.510 ± 0.063. (The true value is $1.10/2.10 \approx 0.524$.)

The model with no covariate estimates the hazard ratio as 0.498 ± 0.064.

The model with the covariate (16) estimates the hazard ratio as 0.001 ± 0.001.

The crude rate ratio looks good, and so does the no-covariate model. However, the model with the covariate says that treatment divides the hazard rate by 1000. Apparently, this is the wrong kind of covariate to put into the model.

If $c = 1$, so that noise offsets the signal in the covariate, the full model estimates a hazard ratio of about 0.45—somewhat too low. If $c = 2$, noise swamps the (bad) signal, and the full model works fine. There is actually a little bit of variance reduction.

Some observers may object that (16) is not a confounder, because (on average) there will be balance between treatment and control. To meet that objection, just change the covariate to

$$Z_i = \exp(-\tau_i) + \zeta_i \exp(-\tau_i/2) + cU_i, \tag{17}$$

where ζ_i is the treatment dummy. The covariate (17) is unbalanced between treatment and control groups. It is related to outcomes. It contains valuable information. In short, it is a classic example of a confounder. But, for the proportional-hazards model, it's the wrong kind of confounder—poison, unless c is quite large.

Here are the results for $c = 2$, when half the variance in (17) is accounted for by noise, so there is a lot of dilution.

The crude rate ratio is 0.522 ± 0.056.

The model with no covariate estimates the hazard ratio as 0.510 ± 0.056.

The model with the covariate (17) estimates the hazard ratio as 0.165 ± 0.138.

(We have independent randomization across examples, which is how 0.510 in the previous example changed to 0.522 here.) Putting the covariate (17) into the model biases the hazard ratio downwards by a factor of 3.

What is wrong with these covariates? The proportional-hazards model is not only about adjusting for confounders, it is also about *hazards that are proportional to the baseline hazard*. The key assumption in the model is something like this. Given that a subject is alive and uncensored at time t, and given the covariate history up to time t, the probability of failure in $(t, t + dt)$ is $h(t)\exp(X_{it}\beta)\,dt$, where h is the baseline hazard rate. In (16) with $c = 0$, the conditional failure time will be known, because Z_i determines τ_i. So the key assumption in the model breaks down. If c is small, the situation is similar, as it is for the covariate in (17).

Some readers may ask whether problems can be averted by judicious use of model diagnostics. No doubt, if we start with a well-defined type of breakdown in modeling assumptions, there are diagnostics that will detect the problem. Conversely, if we fix a suite of diagnostics, there are problems that will evade detection (Freedman, 2008a).

Causal inference from observational data

Freedman (2005) reviews a logical framework, based on Neyman (1923), in which regression can be used to infer causation. There is a straightforward extension to the Cox model with non-stochastic covariates. Beyond the purely statistical assumptions, the chief additional requirement is "invariance to intervention." In brief, manipulating treatment status should not change the statistical relations.

For example, suppose a subject chose the control condition, but we want to know what would have happened if we had put him into treatment. Mechanically, nothing is easier: just switch the treatment dummy from 0 to 1, and compute the hazard rate accordingly. Conceptually, however, we are assuming that the intervention would not have changed the baseline hazard rate, or the values of the other covariates, or the coefficients in the model.

Invariance is a heroic assumption. How could you begin to verify it, without actually doing the experiment and intervening? That is one of the essential difficulties in using models to make causal inferences from non-experimental data.

What is the bottom line?

There needs to be some hard thinking about the choice of covariates, the proportional-hazards assumption, the independence of competing risks, and so forth. In the applied literature, these issues are rarely considered in any depth. That is why the modeling efforts, in observational studies as in experiments, are often unconvincing.

Cox (1972) grappled with the question of what the proportional hazards model was good for. He ends up by saying

> "[i] Of course, the [model] outlined here can be made much more specific by introducing explicit stochastic processes or physical models. The wide variety of possibilities serves to emphasize the difficulty of inferring an underlying mechanism indirectly from failure times alone rather than from direct study of the controlling physical processes. [ii] As a basis for rather empirical data reduction, [the model] seems flexible and satisfactory." [p. 201]

The first point is undoubtedly correct, although it is largely ignored by practitioners. The second point is at best debatable. If the model is wrong, why are the parameter estimates a good summary of the data? In any event, questions about summary statistics seem largely irrelevant: practitioners fit the model to the data without considering assumptions, and leap to causal conclusions.

Where do we go from here?

I will focus on clinical trials. Altman, Schulz, and Moher et al (2001) document persistent failures in the reporting of the data, and make detailed proposals for improvement. The following recommendations are complementary; also see Andersen (1991).

(i) As is usual, measures of balance between the assigned-to-treatment group and the assigned-to-control group should be reported.

(ii) After that should come a simple intention-to-treat analysis, comparing rates (or averages and SDs) among those assigned to treatment and those assigned to the control group.

(iii) Crossover should be discussed, and deviations from protocol.

(iv) Subgroup analyses should be reported, and corrections for crossover if that is to be attempted. Two sorts of corrections are increasingly common. (a) Per-protocol analysis censors subjects who cross over from one arm of the trial to the other, for instance, subjects who are assigned to control but insist on treatment. (b) Analysis by treatment received compares those who receive treatment with those who do not, regardless of assignment. These analyses require special justification (Freedman, 2006).

(v) Regression estimates (including logistic regression and proportional hazards) should be deferred until rates and averages have been presented. If regression estimates differ from simple intention-to-treat results, and reliance is placed on the models, that needs to be explained. The usual models are not justified by randomization, and simpler estimators may be more robust.

(vi) The main assumptions in the models should be discussed. Which ones have been checked. How? Which of the remaining assumptions are thought to be reasonable? Why?

(vii) Authors should distinguish between analyses specified in the trial protocol and other analyses. There is much to be said for looking at the data; but readers need to know how much looking was involved, before that significant difference popped out.

(viii) The exact specification of the models used should be posted on journal websites, including definitions of the variables. The underlying data should be posted too, with adequate documentation. Patient confidentiality would need to be protected, and authors may deserve a grace period after first publication to further explore the data.

Some studies make data available to selected investigators under stringent conditions (Geller et al., 2004), but my recommendation is different. When data-collection efforts are financed by the public, the data should be available for public scrutiny.

Some pointers to the literature

Early publications on vital statistics and life tables include Graunt (1662), Halley (1693), and Bernoulli (1760). Bernoulli's calculations on smallpox may seem a bit mysterious. For discussion, including historical context, see Gani (1978) or Dietz and Hesterbeek (2002). A useful book on the early history of statistics, including life tables, is Hald (2005).

Freedman (2007, 2008b) discusses the use of models to analyze experimental data. In brief, the advice is to do it late if at all. Fremantle et al. (2003) have a critical discussion on use of "composite endpoints," which combine data on many distinct endpoints. An example, not much exaggerated, would be fatal MI + non-fatal MI + angina + heartburn.

Typical presentations of the proportional-hazards model (this one included) involve a lot of hand-waving. It is possible to make math out the hand-waving. But this gets very technical very fast, with martingales, compensators, left-continuous filtrations, and the like. One of the first rigorous treatments was Odd Aalen's Ph. D. thesis at Berkeley, written under the supervision of Lucien LeCam. See Aalen (1978) for the published version, which builds on related work by Pierre Bremaud and Jean Jacod.

Survival analysis is sometimes viewed as a special case of "event history analysis." Standard mathematical references include Andersen et al (1996), Fleming and Harrington (2005). A popular alternative is Kalbfleisch and Prentice (2002). Some readers like Miller (1998); others prefer Lee and Wang (2003). Jewell (2003) is widely used. Technical details in some of these texts may not be in perfect focus. If you want mathematical clarity, Aalen (1978) is still a paper to be recommended.

For a detailed introduction to the subject, look at Andersen and Keiding (2006). This book is organized as a one-volume encyclopedia. Peter Sasieni's entry on the "Cox Regression Model" is a good starting point; after that, just browse. Lawless (2003) is another helpful reference.

Appendix: The delta method in more detail

The context for this discussion is the Women's Health Initiative, a randomized controlled experiment on the effects of hormone replacement therapy. Let N and N' be the numbers of women randomized to treatment and control. Let ξ and ξ' be the corresponding numbers of failures (that is, for instance, fatal or non-fatal heart attacks).

The crude rate ratio is the failure rate in the treatment arm divided by the rate in the control arm, with no adjustments whatsoever. Algebraically, this is $(\xi/N)\big/(\xi'/N')$. The logarithm of the crude rate ratio is

$$\log\xi - \log\xi' - \log N + \log N'. \tag{18}$$

Let $\mu = E(\xi)$. So

$$\begin{aligned}\log\xi &= \log\Big[\mu\Big(1+\frac{\xi-\mu}{\mu}\Big)\Big]\\ &= \log\mu + \log\Big(1+\frac{\xi-\mu}{\mu}\Big)\\ &\approx \log\mu + \frac{\xi-\mu}{\mu},\end{aligned} \tag{19}$$

because $\log(1+h) \approx h$ when h is small. The delta-method $\approx$ a one-term Taylor series.

For present purposes, take ξ to be approximately Poisson, so $\text{var}(\xi) \approx \mu \approx \xi$ and

$$\text{var}\Big(\frac{\xi-\mu}{\mu}\Big) \approx \frac{1}{\mu} \approx \frac{1}{\xi}. \tag{20}$$

A similar calculation can be made for ξ'. Take ξ and ξ' to be approximately independent, so the log of the crude rate ratio has variance approximately equal to $1/\xi + 1/\xi'$.

The modeling is based on the idea that each subject has a small probability of failing during the trial. This probability is modifiable by treatment. Probabilities and effects of treatment may differ from one subject to another. Subjects are assumed to be independent, and calculations are conditional on assignment.

Exact combinatorial calculations can be made, unconditionally, based on the permutations used in the randomization. To take blocking, censoring, or time-to-failure into account, unpublished data would usually be needed.

For additional information on the delta method, see van der Vaart (1998). Many arguments for asymptotic behavior of the MLE turn out to depend on more rigorous (or less rigorous) versions of the delta method. Similar comments apply to the Kaplan-Meier estimator.

References

Aalen, O. O. (1978). "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, 6, 701–26.

Altman, D. G., Schulz, K. F., David Moher et al. (2001). "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration," *Annals of Internal Medicine*, 134, 663–94.

Andersen, P. K. (1991). "Survival Analysis 1982–1991: The Second Decade of the Proportional Hazards Regression Model," *Statistics in Medicine*, 10, 1931–41.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1996). *Statistical Models Based on Counting Processes*. Corr. 4th printing. New York: Springer-Verlag.

Andersen, P. K. and Keiding, N. eds. (2006). *Survival and Event History Analysis*. Chichester, U. K.: John Wiley & Sons.

Barrett-Connor, E. (1991). "Postmenopausal Estrogen and Prevention Bias," *Annals of Internal Medicine*, 115, 455–56.

Bernoulli, D. (1760). "Essai d'une nouvelle analyse de la mortalité causée par la petite variole, et des avantages de l'inoculation pour la prévenir," *Mémoires de Mathématique et de Physique de l'Académie Royale des Sciences*, Paris, 1–45. Reprinted in *Histoire de l'Académie Royale des Sciences* (1766) .

Cox, D. (1972). "Regression Models and Lifetables," *Journal of the Royal Statistical Society*, Series B, 34, 187–220 (with discussion).

Dietz, K. and Heesterbeek, J. A. P. (2002). "Daniel Bernoulli's Epidemiological Model Revisited," *Mathematical Biosciences*, 180, 1–21.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. 2nd rev. edn. New York: John Wiley & Sons.

Freedman, D. A. (1983). *Markov Chains*. New York: Springer-Verlag.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

Freedman, D. A. (2006). "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review*, 30, 691–713.
http://www.stat.berkeley.edu/users/census/oxcauser.pdf

Freedman, D. A. (2007). "On Regression Adjustments in Experiments with Several Treatments," To appear in *Annals of Applied Statistics*.
http://www.stat.berkeley.edu/users/census/neyregcm.pdf

Freedman, D. A. (2008a). "Diagnostics Cannot Have Much Power Against General Alternatives."
http://www.stat.berkeley.edu/users/census/notest.pdf

Freedman, D. A. (2008b). "Randomization Does Not Justify Logistic Regression."
http://www.stat.berkeley.edu/users/census/neylogit.pdf

Freedman, D. A., Purves, R. A., and Pisani R. (2007). *Statistics*. 4th edn. New York: W. W. Norton & Co., Inc.

Fremantle, N., Calvert, M., Wood, J. et al. (2003). "Composite Outcomes in Randomized Trials: Greater Precision But With Greater Uncertainty?" *Journal of the American Medical Association*, 289, 2554–59.

Gani, J. (1978). "Some Problems of Epidemic Theory," *Journal of the Royal Statistical Society*, Series A, 141, 323–47 (with discussion).

Geller, N. L., Sorlie, P., Coady, S., Fleg, J., and Friedman, L. (2004). Limited access data sets from studies funded by the National Heart, Lung, and Blood Institute. *Clinical Trials*, 1, 517–524.

Graunt, J. (1662). *Natural and Political Observations Mentioned in a following Index, and Made upon the Bills of Mortality*. London. Printed by Tho. Roycroft, forJohn Martin, James Allestry, and Tho. Dicas, at the Sign of the Bell in St. Paul's Church-yard, MDCLXII. Available on-line at
http://www.ac.wwu.edu/˜stephan/Graunt/
There is a life table on page 62. Reprinted by Ayer Company Publishers, NH, 2006.

Grodstein, F., Stampfer, M. J., Manson, J. et al. (1996). "Postmenopausal Estrogen and Progestin Use and the Risk of Cardiovascular Disease," *New England Journal of Medicine*, 335, 453–61.

Grodstein, F., Stampfer, M. J., Colditz, G. A. et al. (1997). "Postmenopausal Hormone Therapy and Mortality," *New England Journal of Medicine*, 336, 1769–75.

Hald, A. (2005). *A History of Probability and Statistics and Their Applications before 1750*. New York: John Wiley & Sons.

Halley, E. (1693). "An Estimate of the Mortality of Mankind, Drawn from Curious Tables of the Births and Funerals at the City of Breslaw; with an Attempt to Ascertain the Price of Annuities upon Lives," *Philosophical Transactions of the Royal Society of London*, 196, 596–610, 654–56.

Heiss, G., Wallace, R., Anderson, G. L., et al. (2008). "Health Risks and Benefits 3 Years After Stopping Randomized Treatment With Estrogen and Progestin," *Journal of the American Medical Association*, 299, 1036–45.

Henschke, C. I., Yankelevitz, D. F., Libby, D. M. et al. (2006). "The International Early Lung Cancer Action Program Investigators. Survival of Patients with Stage I Lung Cancer Detected on CT Screening," *New England Journal of Medicine*, 355, 1763–71.

Hill, A. B. (1961). *Principles of Medical Statistics*. 7th ed. London: The Lancet.

Jewell, N. P. (2003). *Statistics for Epidemiology*. Boca Raton, FL: Chapman & Hall/CRC.

Kalbfleisch, J. D. and Prentice, R. L. (1973). "Marginal Likelihoods Based on Cox's Regression and Life Model," *Biometrika*, 60, 267–78.

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

Kaplan, E. L. and Meier, P. (1958). "Nonparametric Estimation from Incomplete Observations," *Journal of American Statistical Association*, 53, 457–81.

Kirk, D. (1996). "Demographic Transition Theory," *Population Studies*, 50, 361–87.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. 2nd ed. New York: John Wiley & Sons.

Lee, E. T. and Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*. 3rd ed. New York: John Wiley & Sons.

Miller, R. G., Jr. (1998). *Survival Analysis*. New York: John Wiley & Sons.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10: 1–51, in Polish. English translation by Dabrowska, D. M. and Speed, T. P. (1990), *Statistical Science*, 5, 465–80 (with discussion).

Pargament, K. I., Koenig, H. G., Tarakeshwar, N., and Hahn, J. (2001). "Religious Struggle as a Predictor of Mortality among Medically Ill Patients," *Archives of Internal Medicine*, 161, 1881–85.

Patz, E. F., Jr., Goodman, P. C., and Bepler, G. (2000). "Screening for Lung Cancer," *New England Journal of Medicine*, 343, 1627–33.

Petitti, D. B. (1994). "Coronary Heart Disease and Estrogen Replacement Therapy: Can Compliance Bias Explain the Results of Observational Studies?" *Annals of Epidemiology*, 4, 115–18.

Petitti, D. B. (1998). "Hormone Replacement Therapy and Heart Disease Prevention: Experimentation Trumps Observation," *Journal of the American Medical Association*, 280, 650–51.

Petitti, D. B. (2002). "Hormone Replacement Therapy for Prevention," *Journal of the American Medical Association*, 288, 99–101.

Petitti, D. B. and Freedman, D. A. (2005). "Invited Commentary: How Far Can Epidemiologists Get with Statistical Adjustment?" *American Journal of Epidemiology*, 162, 415–18.

Petitti, D. B. and Chen, W. (2008). "Statistical Adjustment for a Measure of Healthy Lifestyle Doesn't Yield the Truth about Hormone Therapy," to appear in *Probability and Statistics: Essays in Honor of David A. Freedman*, eds. Deborah Nolan and Terry Speed. Institute of Mathematical Statistics.

Rossouw, J. E., Anderson, G. L., Prentice, R. L. et al. (2002). "Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women'S Health Initiative Randomized Controlled Trial," *Journal of the American Medical Association*, 288, 321–333.

Rudin, W. (1976). *Principles of Mathematical Analysis*. 3rd. ed. New York: McGraw-Hill.

Shapiro, S., Venet, W., Strax, P., and Venet, L. (1988). *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986*. Baltimore: Johns Hopkins University Press.

Thiébaut, A. C. M. and Bénichou, J. (2004). "Choice of time-scale in Cox's model analysis of epidemiologic cohort data: A simulation study," *Statistics in Medicine*, 23, 3803–3820.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge, U. K.: Cambridge University Press.

Welch, H. G., Woloshin, S., Schwartz, L. M. et al. (2007). "Overstating the Evidence for Lung Cancer Screening: The International Early Lung Cancer Action Program (I-ELCAP) Study," *Archives of Internal Medicine*, 167, 2289–95.

Key words and phrases

Survival analysis, event history analysis, life tables, Kaplan-Meier estimator, proportional hazards, Cox model.

Author's footnote

David A. Freedman is Professor of Statistics, University of California, Berkeley CA 94720-3860 (E-mail: freedman@stat.berkeley.edu). Charles Kooperberg, Russ Lyons, Diana Petitti, Peter Sasieni, and Peter Westfall were very helpful. Kenneth Pargament generously answered questions about his study.