A Systems Approach to Epidemiology

Aaron Kershenbaum

Research done while at

Columbia University School of Public Health

Epidemiology

The study of disease

- Causes
- Spread
- Outcomes
- Treatments
- Public health

Covariates

Factors that affect the risk of getting a disease and disease outcomes

Age

Gender

Socioeconomic status

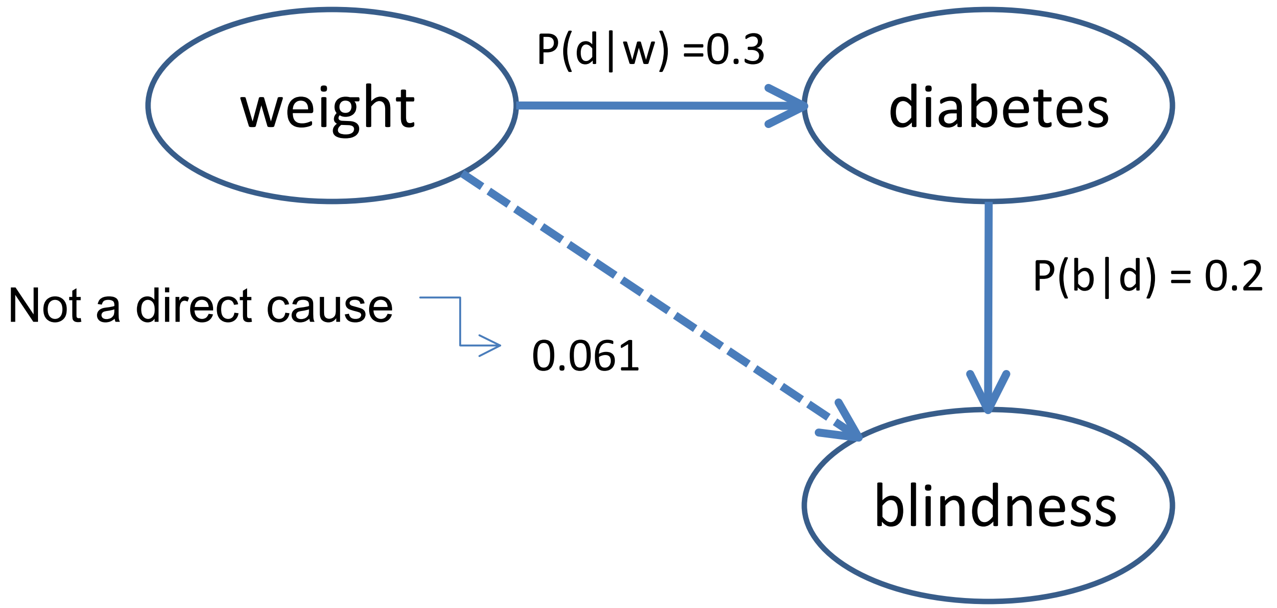
Race

Genome

Weight

…

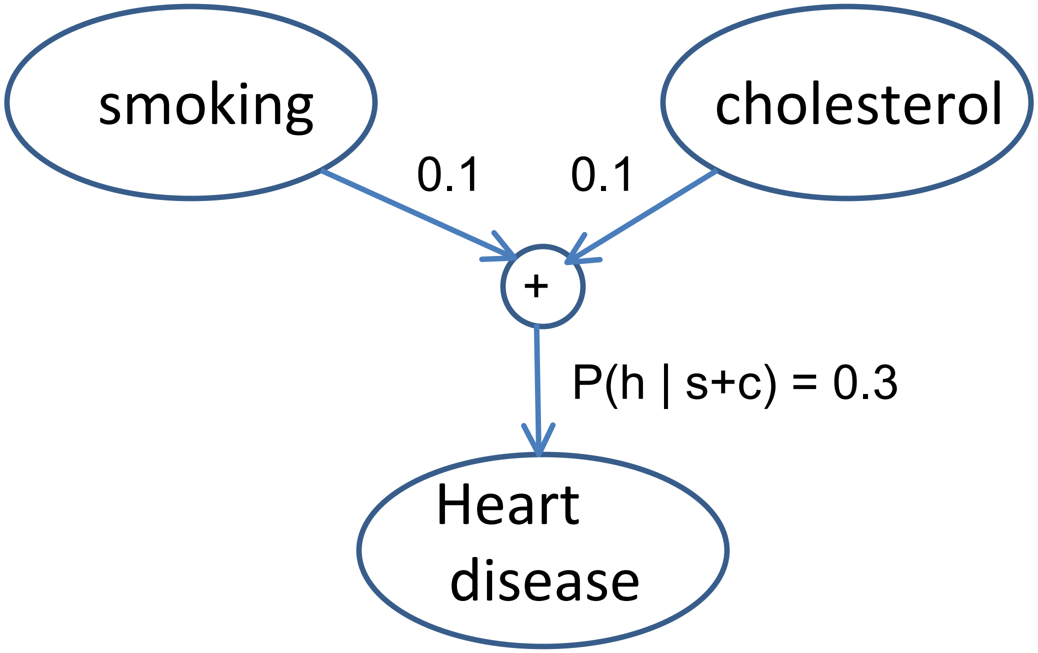
Covariates are interrelated



The interrelated covariates can be analyzed as a system.

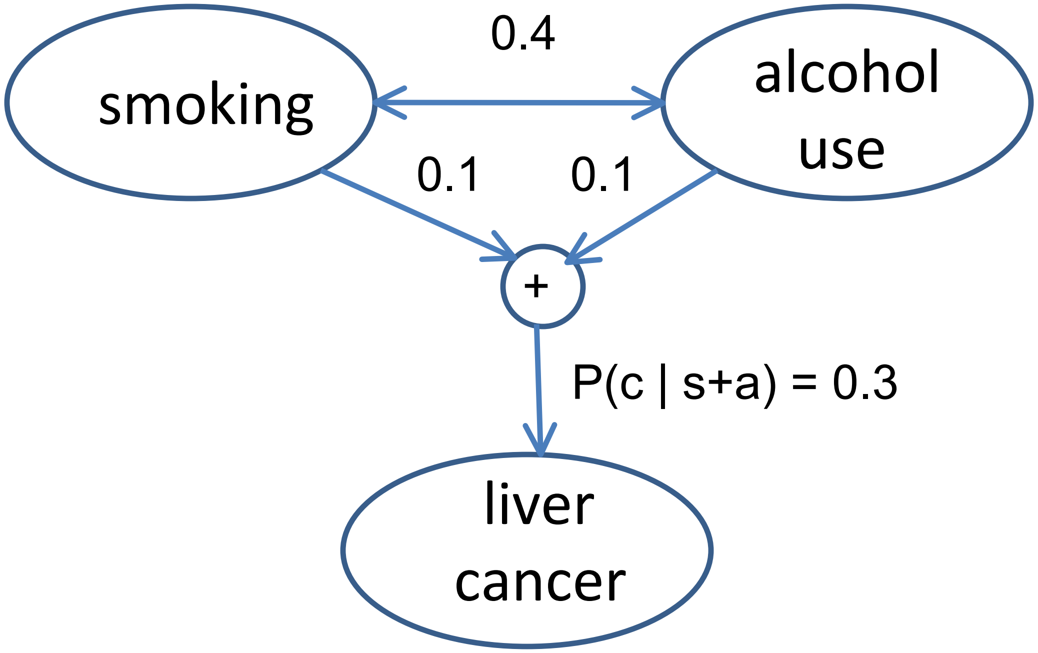
Identify what is: a direct cause, an indirect cause, not a cause

Covariates are interrelated



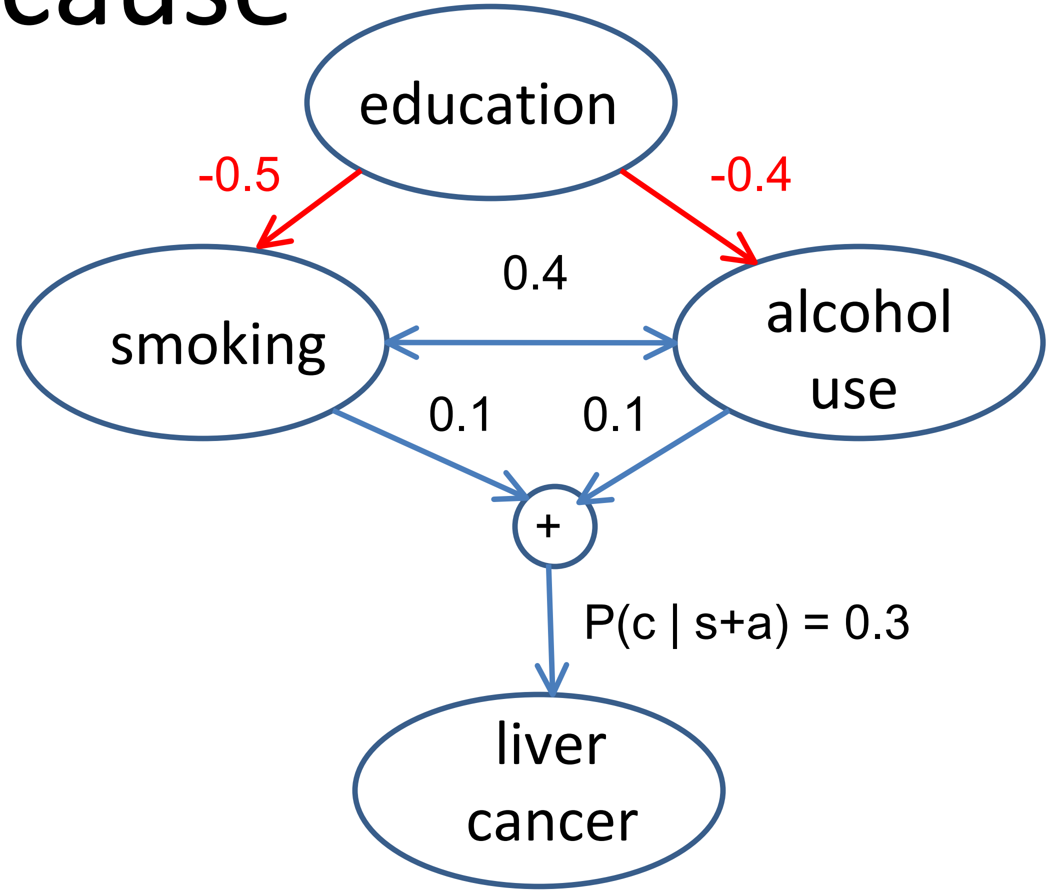
The interrelated covariates can reinforce one another or attenuate one another

Covariates are interrelated



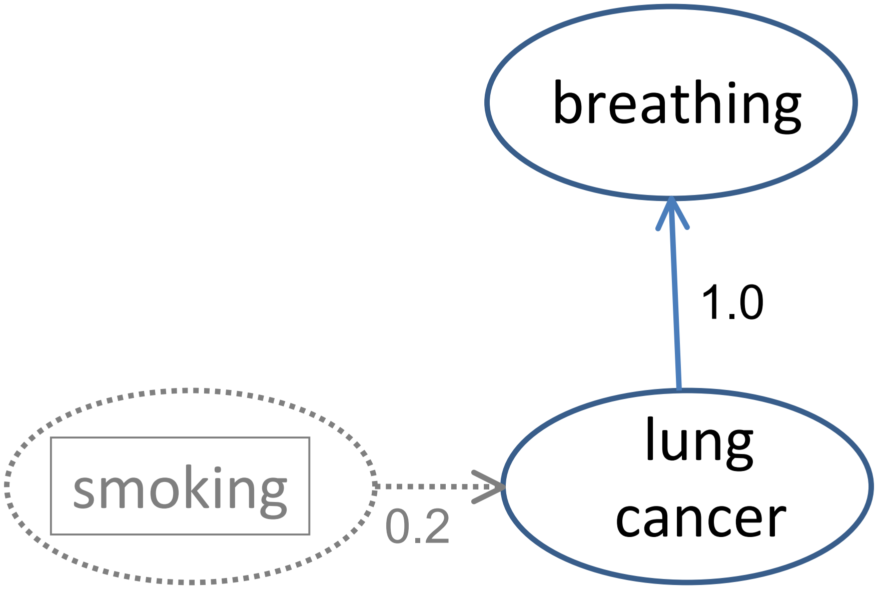
Covariates can themselves be correlated

Common cause



Risk can be reduced by eliminating a common cause even though the cause is not itself directly associated with the disease.

Cause and Effect



One might conclude that lung cancer causes breathing.

Correlation does not imply causality.

Real causes may not even be part of the analysis.

Analysis in Epidemiology

Lot of challenges:

- Data is hard to obtain
- Causes are interrelated
- Causes are easily confused with one another and with effects

Solution approach:

- Careful design of studies
- Including medical knowledge
- Appropriate analytic methods

Study Design: Case / control

Select cases and controls matched on

as many covariates as possible

Follow them over time using

Direct observation

Medical records

Pros and cons:

Effect of covariates can be managed

Additional data can be collected

Expensive → smaller studies

Study Design: Retrospective

Select a cohort (cases and controls)

Analyze existing data (e.g., medical records)
for members of the cohort

Pros and cons:

Can work with large cohorts (*)

Can look over a longer timeframe (*)

Less control over cohort

→ It is easier to introduce bias

(*) If appropriate data exists

Goal: Impact on public health

→Deal with an important problem

→Cover of a significant part of the population

Large support in cohort

Broadly applicable

→Convincing results

Clear outcome

Statistically significant

→Accurate results

The impact should be positive !

Analytic Approach

	#Case	#Ctrl
# Exposed	A	B
# Not exposed	C	D

METRICS:

(*) Risk $= A / (A + B)$

Relative Risk $= [(A/(A+B)] / [(C/(C+D)]$

Odds Ratio $= (A/B) / (C/D) = AD / BC$

(*) Only requires data from exposed people

Generalizability

	#Case	#Ctrl
# Exposed	A	B
# Not exposed	C	D

Results can only be generalized to the part of the population for which the cohort is representative

This in turn requires either a very large cohort or a willingness to accept less control over the cohort composition with the attendant risk of bias.

Confidence in Results

	#Case	#Ctrl
# Exposed	A	B
# Not exposed	C	D

OR = AD / BC

If we assume that OR is a normally distributed random variable (not always a valid assumption), we can compute a confidence interval using standard statistical tests; e.g., Chi squared, T-test based on the properties of the normal distribution.

Ex. OR = 2.21 ; CI 95% [1.98,2.49]

Fisher’s Exact Test

	#Case	#Ctrl	
# Exposed	A	B	A+B
# Not exposed	C	D	C+D
	A+C	B+D	N

OR = 2.21

pValue = Prob {OR has a value >= 2.21 or more by chance}

If we are willing to assume that all possible arrangements of A, B, C, and D are equally likely, then we can count the number of ways A, B, C, and D can take values yielding OR >= 2.21, subject to the restrictions implied by the row and column sums.

Fisher’s Exact Test

	#Case	#Ctrl	
# Exposed	A	B	A+B
# Not exposed	C	D	C+D
	A+C	B+D	N

OR = 2.21

At first glance, this seems to require a huge amount of computation.
But in fact the value of A fixes the values of B, C and D.
Futhermore, N(A,B C,D), the number of ways A, B, C and D can occur is:
N(A,B C,D)= [(A+B)! (C+D)! (A+C)! (B+D)!] / [A! B! C! D! (A+B+C+D)!]
and
pValue = SUM on A = 0 , … min(0 , min(A+B),(A+C)) of N(A,B C,D)
We avoid numerical problems by storing values of logarithms of factorials

Survival Models

In order to quantitatively assess and coherently explain the effect of an exposure (e.g., treatment) on survival, we need a functional model of survival.

The simplest model is to say that the probability of survival over a small interval of time, t, is constant but the constant is different for people who are exposed (s_1) in comparison with unexposed people (s_2).

If T = nt (n>>1) then the probability of person surviving past T is $(s_1)^n$, for exposed people and $(s_2)^n$, for unexposed people.

Since $e = \text{Lim}_{n \rightarrow \text{INF}} (1 - 1/n)^n$

For large n, these values converge to exp($-n^*f_1$) and exp($-n^*f_2$)
where $f_i = 1 – s_i$
and the effect of treatment on survival is the ratio of these two quantities

Survival Models

Of course the assumption that the instantaneous probability of survival remains constant over a long period of time may not be true. The assumption must be tested. The simplest way of doing this is to plot the log of survival as a function of time and see how close the plot comes to a straight line. Here are also numerical tests that are best carried out with a computer.

If the assumption holds true over shorter time intervals it is possible to analyze each interval separately, but the explanation becomes more complex as the number of subintervals increases.

Survival Models

A more general assumption is that the probability of survival as a function of time is of the form

$$p_s(t) = f(t) * \exp(-s)$$

This, of course, is totally general since f(t) could be anything.

But now suppose we can model the survival probabilities with and without treatment as

$$p_{s1}(t) = f(t) * \exp(-s_1) \text{ and } p_{s2}(t) = f(t) * \exp(-s_2)$$

Their ratio is then * exp ($-(s_1-s_2)$); i.e., their ratio is a constant

This is known as a proportional hazards model

Such models are important because they can be generalized further to take into account the effects of covariates, compute these effects, and most importantly, describe these effects in a simple way.

But, again, this relies on the assumption that the survival function is of this form

Cox Proportional Hazards Model

Suppose we want to find the relationship of survival to an exposure, say a particular medication or compare the effects of two medications in the presence of covariates; e.g., age (a), gender (g), race (r), and disease stage (d), each of which can take different values from a finite set. The survival function is now generalized to

$$p_s(t) = f(t) * \exp \{ - [c_a a + c_g g + c_r r + c_d d] \}$$

We can also define, T, the average survival time as the expected value of t given the probability density function $p_s(t)$.

We have thus succeeded in building a multiplicative model including all the covariates and relating their effects on survival time, T.

The multiplicative model is consistent with the assumption (which may not be true) that the effect of the covariates is to alter the instantaneous survival rate.

Note that if we divide value of $p_s(t)$ for two different values of a covariate, we get a constant; hence the name proportional hazard model.

Cox Proportional Hazards Model

If we consider the log of $p_s(t)$ we have

$\log [p_s(t)] = \log[f(t)] + c_a*(-a) + c_g*(-g) + c_r*(-r) + c_d*(-d)$

And we can determine the values of the parameters, c_i using linear regression, a well-known method for fitting data to a linear model, minimizing the sum of the squares of the error between observed values and the value given by the model.

We now have a model that relates the values of the covariates to survival in a clear and simple way; i.e., we have a quantitative model for the system model formed by the covariates.

Given such a clear model we now have hopes of actually affecting public health policy.

Cox Proportional Hazards Model

The model

$$p_s(t) = f(t) * \exp \{ - [c_a a + c_g g + c_r r + c_d d] \}$$

assumes at the effects of the covariates are independent of one another. This is not always. Consider the effects of smoking and cholesterol on heart disease.

We can extend the model above by adding variables corresponding to combinations of covariates; e.g., $c_{ag}ag$.

This can be done manually or combinatorially.

This is particularly attractive when considering genetic factors where the effect is likely to be from a combination of genes that individually show little effect. When the number of genes being considered is large, the combinatorics is daunting.

Genetic Models

The notion of considering combinations of factors is particularly attractive when considering genetic factors where the effect is likely to be from a combination of genes that individually show little effect.

Genetic models can be formed based on models relating the function of proteins that are synthesized by specific genes, more specifically, the variations in alleles (single nucleotides) within the genes. These are called single nucleotide polymorphisms (SNPs). This is another type of systems model.

When the number of SNPs being considered is large, the combinatorics is daunting. Over 3 million SNPs have already been identified in the human genome and one can test for one million of them simultaneously using microchips which now cost roughly $400. $16^{3000000}$ is a very large number!

Genetic models do not require analysis of clinical data, although ultimately the conclusions reached using them requires validation using clinical data.

Genetic modeling forms a separate branch of epidemiology.

The Kaplan-Meier Procedure

In most real epidemiological studies we also have to deal with the problem of that some people leave the study before it ends. . This is called censoring.

If what we are calling survival time) is not time to death(e.g., time to onset of a disease, people may leave because they die.

We also remain concerned about all the assumptions we have been making about the functional form of the survival function.

The Kaplan-Meier procedure deals with both of these problems. Its drawback is that it yields a survival curve rather than a function and it only compares different values for a single variable before becoming unwieldy. (We must still remember that our goal is to produce results that others are willing to act on.)

The Kaplan-Meier Procedure

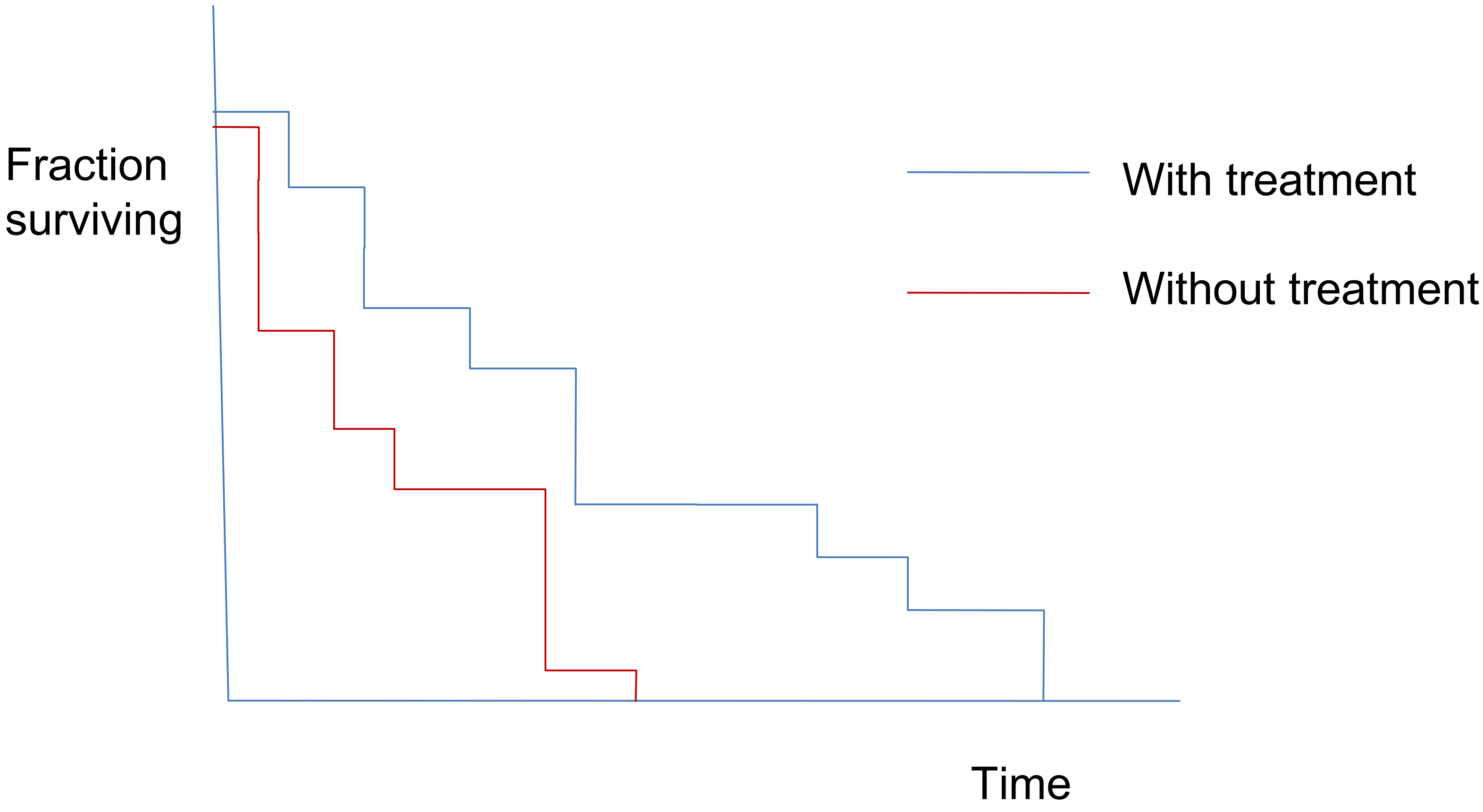
The Kaplan-Meier procedure is a product-limit calculation which gives the fraction of people surviving to the end of interval A_n as the product of the fractions of people surviving through all intervals up to that point; i.e.,

$$S_n = s_1 * s_2 *\dots s_n$$

Note that this is similar to the functional forms we used before but we now make no assumptions about the survival rate remaining constant beyond the length of a single interval, which we can make as short as the time between “deaths”.

The s_i are computed directly as the fraction of people at risk that survive to the end of interval A_i . The definition of at risk is still alive and still in the study. As people leave the study through death or censoring, they are no longer part of the denominator of the fraction.

The Kaplan-Meier Curve



Extensions and Conclusions

The proportional hazards model can be extended to account for censoring.

The Kaplan-Meier procedure can be extended to account for covariates.

The general approaches described here can be used with other functional forms.

Simplicity can be traded for accuracy.

Larger cohorts allow for greater confidence.

Biases may be entirely missed.

Causes may be missed or confused with effect.

Interactions may be missed.

Ultimately, mathematical results must be reconciled with medical knowledge.