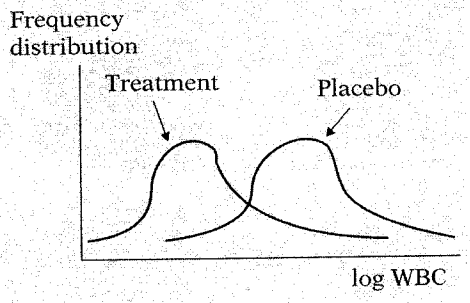
## EXAMPLE: CONFOUNDING

Treatment group: $\overline{\log \text{WBC}} = 1.8$

Placebo group: $\overline{\log \text{WBC}} = 4.1$

Indicates **confounding** of treatment effect by log WBC



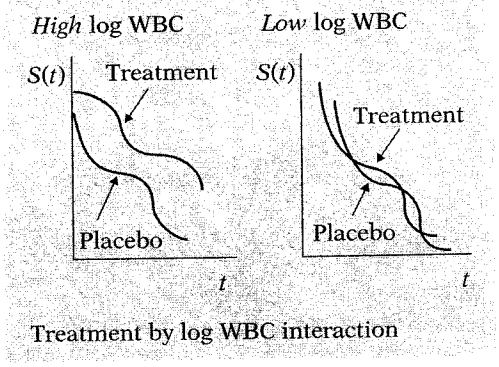

Need to adjust for imbalance in the distribution of log WBC

Although a full exposition of the nature of confounding is not intended here, we provide a simple scenario to give you the basic idea. Suppose all of the subjects in the treatment group had very low log WBC, with an average, for example, of 1.8, whereas all of the subjects in the placebo group had very high log WBC, with an average of 4.1. We would have to conclude that the results we've seen so far that compare treatment with placebo groups may be misleading.

The additional information on log WBC would suggest that the treatment group is surviving longer simply because of their low WBC and not because of the efficacy of the treatment itself. In this case, we would say that **the treatment effect is confounded by the effect of log WBC.**

More typically, the distribution of log WBC may be quite different in the treatment group than in the control group. We have illustrated one extreme in the graph at the left. Even though such an extreme is not likely, and is not true for the data given here, the point is that some attempt needs to be made to adjust for whatever imbalance there is in the distribution of log WBC.

## EXAMPLE: INTERACTION




Treatment by log WBC interaction

Another issue to consider regarding the effect of log WBC is **interaction.** What we mean by interaction is that the effect of the treatment may be different, depending on the level of log WBC. For example, suppose that for persons with high log WBC, survival probabilities for the treatment are consistently higher over time than for the placebo. This circumstance is illustrated by the first graph at the left. In contrast, the second graph, which considers only persons with low log WBC, shows no difference in treatment and placebo effect over time. In such a situation, we would say that **there is strong treatment by log WBC interaction,** and we would have to qualify the effect of the treatment as depending on the level of log WBC.

Need to consider:

- interaction;
- confounding.

**The problem:**

Compare two groups after adjusting for confounding and interaction.

## EXAMPLE

| | Individual # | $t$ (weeks) | $\delta$ | $X_1$ (Group) | $X_2$ (log WBC) |
|---|---|---|---|---|---|
| Group 1 | 1 | 6 | 1 | 1 | 2.31 |
| | 2 | 6 | 1 | 1 | 4.06 |
| | 3 | 6 | 1 | 1 | 3.28 |
| | 4 | 7 | 1 | 1 | 4.43 |
| | 5 | 10 | 1 | 1 | 2.96 |
| | 6 | 13 | 1 | 1 | 2.88 |
| | 7 | 16 | 1 | 1 | 3.60 |
| | 8 | 22 | 1 | 1 | 2.32 |
| | 9 | 23 | 1 | 1 | 2.57 |
| | 10 | 6 | 0 | 1 | 3.20 |
| | 11 | 9 | 0 | 1 | 2.80 |
| | 12 | 10 | 0 | 1 | 2.70 |
| | 13 | 11 | 0 | 1 | 2.60 |
| | 14 | 17 | 0 | 1 | 2.16 |
| | 15 | 19 | 0 | 1 | 2.05 |
| | 16 | 20 | 0 | 1 | 2.01 |
| | 17 | 25 | 0 | 1 | 1.78 |
| | 18 | 32 | 0 | 1 | 2.20 |
| | 19 | 32 | 0 | 1 | 2.53 |
| | 20 | 34 | 0 | 1 | 1.47 |
| | 21 | 35 | 0 | 1 | 1.45 |

The example of interaction we just gave is but one way interaction can occur; on the other hand, interaction may not occur at all. As with confounding, it is beyond our scope to provide a thorough discussion of interaction. In any case, the assessment of interaction is something to consider in one's analysis in addition to confounding that involves explanatory variables.

Thus, with our extended data example, the basic **problem** can be described as follows: to compare the survival experience of the two groups after adjusting for the possible confounding and/or interaction effects of log WBC.

The problem statement tells us that we are now considering two explanatory variables in our extended example, whereas we previously considered the single variable, group status. The data layout for the computer needs to reflect the addition of the second variable, log WBC. The extended table in computer layout form is given at the left. Notice that we have labeled the two explanatory variables $X_1$ (for group status) and $X_2$ (for log WBC). The variable $X_1$ is our primary study or exposure variable of interest here, and the variable $X_2$ is an extraneous variable that we are interested in adjusting for because of either confounding or interaction.

**EXAMPLE (continued)**

| | Individual # | $t$ (weeks) | $\delta$ | $X_1$ (Group) | $X_2$ (log WBC) |
|---|---|---|---|---|---|
| Group 2 | 22 | 1 | 1 | 0 | 2.80 |
| | 23 | 1 | 1 | 0 | 5.00 |
| | 24 | 2 | 1 | 0 | 4.91 |
| | 25 | 2 | 1 | 0 | 4.48 |
| | 26 | 3 | 1 | 0 | 4.01 |
| | 27 | 4 | 1 | 0 | 4.36 |
| | 28 | 4 | 1 | 0 | 2.42 |
| | 29 | 5 | 1 | 0 | 3.49 |
| | 30 | 5 | 1 | 0 | 3.97 |
| | 31 | 8 | 1 | 0 | 3.52 |
| | 32 | 8 | 1 | 0 | 3.05 |
| | 33 | 8 | 1 | 0 | 2.32 |
| | 34 | 8 | 1 | 0 | 3.26 |
| | 35 | 11 | 1 | 0 | 3.49 |
| | 36 | 11 | 1 | 0 | 2.12 |
| | 37 | 12 | 1 | 0 | 1.50 |
| | 38 | 12 | 1 | 0 | 3.06 |
| | 39 | 15 | 1 | 0 | 2.30 |
| | 40 | 17 | 1 | 0 | 2.95 |
| | 41 | 22 | 1 | 0 | 2.73 |
| | 42 | 23 | 1 | 0 | 1.97 |

Analysis alternatives:

- stratify on log WBC;
- use math modeling, e.g., proportional hazards model.

As implied by our extended example, which considers the possible confounding or interaction effect of log WBC, we need to consider methods for adjusting for log WBC and/or assessing its effect in addition to assessing the effect of treatment group. The two most popular alternatives for analysis are the following:

- to stratify on log WBC and compare survival curves for different strata; or
- to use mathematical modeling procedures such as the proportional hazards or other survival models; such methods will be described in subsequent chapters.

## IX. Multivariable Example

- Describes general multivariable survival problem.
- Gives analogy to regression problems.

We now consider one other example. Our purpose here is to describe a more general type of multivariable survival analysis problem. The reader may see the analogy of this example to multiple regression or even logistic regression data problems.

## EXAMPLE

13-year follow-up of fixed cohort from Evans County, Georgia

$n$ = 170 white males (60+)

$T$ = years until death
Event = death

Explanatory variables:
- exposure variable
- confounders
- interaction variables

*Exposure:*
Social Network Index **(SNI)**


Absence of social network — Excellent social network

*Study goal:* to determine whether **SNI** is protective against death,
i.e., **SNI** $\nearrow \Rightarrow S(t) \nearrow$

*Explanatory variables:*

**SNI** — Exposure variable

**AGE**, **SBP**, **CHR**, **QUET**, **SOCL** — Potential confounders/ interaction variables

Note: $\text{QUET} = \dfrac{\text{weight}}{(\text{height})^2} \times 100$

We consider a data set developed from a 13-year follow-up study of a fixed cohort of persons in Evans County, Georgia, during the period 1967–1980 (Schoenbach et al., *Amer. J. Epid.*, 1986). From this data set, we focus on a portion containing $n$ = 170 white males who are age 60 or older at the start of follow-up in 1967.

For this data set, the outcome variable is $T$, time in years until death from start of follow-up, so the event of interest is **death.** Several explanatory variables are measured, one of which is considered the primary exposure variable; the other variables are considered as potential confounders and/or interaction variables.

The primary exposure variable is a measure called Social Network Index (SNI). This is an ordinal variable derived from questionnaire measurement and is designed to assess the extent to which a study subject has social contacts of various types. With the questionnaire, a scale is used with values ranging from 0 (absence of any social network) to 5 (excellent social network).

The study's goal is to determine whether one's social network, as measured by SNI, is protective against death. If this study hypothesis is correct, then the higher the social network score, the longer will be one's survival time.

In evaluating this problem, several explanatory variables, in addition to SNI, are measured at the start of follow-up. These include AGE, systolic blood pressure (SBP), an indicator of the presence or absence of some chronic disease (CHR), body size as measured by Quetelet's index (QUET = weight over height squared times 100), and social class (SOCL).

These five additional variables are of interest because they are thought to have their own special or collective influence on how long a person will survive. Consequently, these variables are viewed as potential confounders and/or interaction variables in evaluating the effect of social network on time to death.

## EXAMPLE (continued)

**The problem:**
To describe the relationship between **SNI** and time to death, after controlling for **AGE, SBP, CHR, QUET,** and **SOCL.**

*Goals:*

- Measure of effect (adjusted)
- Survivor curves for different SNI categories (adjusted)
- Decide on variables to be adjusted
- Determine method of adjustment

Computer layout: 13-year follow-up study (1967–1980) of a fixed cohort of $n = 170$ white males (60+) from Evans County, Georgia

| # | $t$ | $\delta$ | SNI | AGE | SBP | CHR | QUET | SOCL |
|---|---|---|---|---|---|---|---|---|
| 1 | $t_1$ | $\delta_1$ | $SNI_1$ | $AGE_1$ | $SBP_1$ | $CHR_1$ | $QUET_1$ | $SOCL_1$ |
| 2 | $t_2$ | $\delta_2$ | $SNI_2$ | $AGE_2$ | $SBP_2$ | $CHR_2$ | $QUET_2$ | $SOCL_2$ |
| 170 | $t_{170}$ | $\delta_{170}$ | $SNI_{170}$ | $AGE_{170}$ | $SBP_{170}$ | $CHR_{170}$ | $QUET_{170}$ | $SOCL_{170}$ |

We can now clearly state the problem being addressed by this study: To describe the relationship between SNI and time to death, controlling for AGE, SBP, CHR, QUET, and SOCL.

Our goals in using survival analysis to solve this problem are as follows:

- to obtain some measure of effect that will describe the relationship between SNI and time until death, after adjusting for the other variables we have identified;
- to develop survival curves that describe the probability of survival over time for different categories of social networks; in particular, we wish to compare the survival of persons with excellent networks to the survival of persons with poor networks. Such survival curves need to be adjusted for the effects of other variables.
- to achieve these goals, two intermediary goals are to decide which of the additional variables being considered need to be adjusted and to determine an appropriate method of adjustment.

The computer data layout for this problem is given at the left. The first column lists the 170 individuals in the data set. The second column lists the survival times, and the third column lists failure or censored status. The remainder of the columns list the 6 explanatory variables of interest, starting with the exposure variable SNI and continuing with the variables to be adjusted in the analysis.

## X. Math Models in Survival Analysis

General framework

$E$ $D$

Controlling for $C_1, C_2, \ldots C_p$,

**SNI** study:

**$E$ = SNI ➡ $D$ = survival time**

Controlling for **AGE, SBP, CHR, QUET,** and **SOCL**

| | Model | Outcome |
|---|---|---|
| | Survival | Time to event (with censoring) |
| follow-up time info not used | Linear regression | Continuous (SBP) |
| follow-up time info not used | Logistic regression | Dichotomous (CHD yes/no) |

It is beyond the scope of this presentation to provide specific details of the survival analysis of these data. Nevertheless, the problem addressed by these data is closely analogous to the typical multivariable problem addressed by linear and logistic regression modeling. Regardless of which modeling approach is chosen, the typical problem concerns describing the relationship between an exposure variable (e.g., $E$) and an outcome variable (e.g., $D$) after controlling for the possible confounding and interaction effects of additional variables (e.g., $C_1$, $C_2$, and so on up to $C_p$). In our survival analysis example, $E$ is the social network variable SNI, $D$ is the survival time variable, and there are $p = 5$ $C$ variables, namely, AGE, SBP, CHR, QUET, and SOCL.

Nevertheless, an important distinction among modeling methods is the type of outcome variable being used. In survival analysis, the outcome variable is "time to an event," the event being death, and there is censored data. In linear regression modeling, the outcome variable is generally a continuous variable, like blood pressure. In logistic modeling, the outcome variable is a dichotomous variable, like CHD status, yes or no. And with linear or logistic modeling, we usually do not have information on follow-up time available.

As with linear and logistic modeling, one statistical goal of a survival analysis is to obtain some measure of effect that describes the exposure–outcome relationship adjusted for relevant extraneous variables.

**Measure of effect:**

Linear regression:
regression coefficient β

In linear regression modeling, the measure of effect is usually some regression coefficient β.

Logistic regression:
odds ratio $e^{\beta}$

In logistic modeling, the measure of effect is an odds ratio expressed in terms of an exponential of one or more regression coefficients in the model, for example, $e$ to the β.

Survival analysis:
hazard ratio $e^{\beta}$

In survival analysis, the measure of effect obtained is called a **hazard ratio;** as with the logistic model, this hazard ratio is expressed in terms of an exponential of a regression coefficient in the model.