Tutorial: Survival Analysis— A Statistic for Clinical, Efficacy, and Theoretical Applications

Frederic A. Gruber*
University of Wisconsin–Madison

Current demands for increased research attention to therapeutic efficacy, efficiency, and also for improved developmental models call for analysis of longitudinal outcome data. Statistical treatment of longitudinal speech and language data is difficult, but there is a family of statistical techniques in common use in medicine, actuarial science, manufacturing, and sociology that has not been used in speech or language research. Survival analysis is introduced as a method that avoids many of the statistical problems of other techniques because it treats time as the outcome. In survival analysis, probabilities are calculated not just for groups but also for individuals in a group. This is a major advantage for clinical work. This paper provides a basic introduction to nonparametric and semiparametric survival analysis using speech outcomes as examples. A brief discussion of potential conflicts between actuarial analysis and clinical intuition is also provided.

KEY WORDS: survival analysis, efficacy, prediction, phonology, language

There is a family of statistical methods designed for the analysis of longitudinal data that has potential usefulness in studying both normal and disordered speech and language development. In the social sciences, it is called "event history analysis," whereas in engineering and manufacturing, these methods are called "reliability" or "failure" analysis. The term employed here, "survival analysis," follows the convention of the medical professions.

This paper provides a general introduction to survival analysis. It reviews why survival analysis avoids some of the pitfalls common to more familiar statistical methods, has advantages for clinical speech and language applications, and is timely for the study of treatment efficacy and efficiency. A tutorial on techniques of most likely application to speech and language disorders is presented. Finally, a scenario is offered illustrating how these techniques might support or conflict with clinical intuition.

Most research concerning individual behavior change treats time or age as a predictor (Willett & Singer, 1989, 1991). Age or time is treated as an independent variable. A measure of behavior is treated as the dependent variable. In survival analysis, *time is treated as the outcome.* Instead of asking how rapidly clients change over time, the researcher asks, "How much time must pass before a client displays a change in

*Currently affiliated with Lamar University, Beaumont, TX.

1092-4388/99/4202-0432

X?" Because time is considered an outcome, it is not considered causative.

A second type of question addressed in survival analysis is "Given that a client has not changed by time period *t*, what are the chances the client will change by the next time period(s)?" In communicative disorders, the question addressed might be "How long will it be before X speaks normally?" or "...pronounces /s/ correctly?" or "Do you think X will speak normally by this time next year?"

A third type of question that can be addressed in survival analysis is "How do probable outcome times differ between groups of clients?" or "How are different treatments likely to influence outcome times?" These questions correspond to "Do children with chronic otitis media take longer to reach the same level of speaking proficiency than children who do not have chronic otitis media?" and "Which therapy regimen will most likely achieve a therapy goal more quickly?"

Survival analytic techniques are designed to establish empirical estimates of the probabilities that an individual will experience a qualitative *change of state* at a given time. Change of state may be defined as the crossing of some consistent threshold. Such thresholds might include well-defined therapy goals such as 90% of consonants correct in conversational speech, no disfluencies in 3 minutes of conversation with a stranger, or the correct use of 10 successive prepositions in narrative discourse.

Some Statistical Problems in Longitudinal Studies

Many difficulties intrinsic to the more familiar and commonly used statistical treatments for longitudinal data such as correlation, regression analysis, growth curves, structural equation modeling, and analysis of variance can be avoided by adopting survival analysis. These difficulties include ceiling and floor effects, heterogeneity of variance, treating language as a fixed effect, autocorrelation, and regression to the mean (Blossfeld & Rohwer, 1995; Clark, 1973; Gruber, 1997).

Ceiling and Floor Effects and Heterogeneity of Variance

Ceiling and floor effects occur when a task is either very easy or extremely difficult for participants. The distribution of scores on a task that is too easy pile up on the "correct" side (i.e., negatively skewed), with little variance and high mean correct scores. Tasks that are too difficult show a positively skewed distribution, low mean correct scores, and, again, little variance. Heterogeneous participant samples appear homogenous when judged by measures near floors and ceilings. All of the traditional statistical methods, parametric or nonparametric, then become invalid (Campbell & Boruch, 1975; Hayes, 1973; Marascuilo & Levin, 1983).

Survival analysis avoids ceiling and floor effects because the single score used is the time (age) when a participant achieved a score that crossed the chosen threshold. As long as a high proportion of participants in a study cross the threshold during a study, there are no problems with heterogeneity and ceiling and floor effects.

Language as a Fixed Effect and Autocorrelation

Almost all speech and language studies treat speech and language as a fixed effect. Usually only a few relatively small speech or language samples are taken. For example, articulation, vocabulary, and many syntactic tests use the same small set of words or sentences to gauge the overall language performance of participants. These speech or language samples are incorrectly treated as though they represented the whole linguistic repertoire of the language, leading to considerable measurement error (Clark, 1973; Leonard & Orchard, 1996).

Furthermore, in most clinical studies the number of participants is small compared to the population size, resulting in sampling error. When this source of error combines with measurement and other experimental error, the primary error term in statistical equations can become large. Thus, study results could well misrepresent the real-life situation (Nunnally, 1978).

If these error terms are correlated over time, they are not independent. This problem is described as "autocorrelation," and "serial dependency." This is a serious violation of a basic assumption of parametric statistics (Kirk, 1982; Maxwell & Delaney, 1990). Sources of such bias include the failure to use uninformed, external evaluators of speech or language status, the use of reliability training and consensus transcription techniques, and the use of forced-choice measures, which constrain the kinds of errors possible. Thus, some of the errors made in collecting and scoring data are consistently biased across measurements. Parametric statistics are robust, with some measurement error, but only if the errors are randomly distributed.

In survival analysis, autocorrelation does not create a statistical calculation problem, again, because only a single measurement for each participant enters into the calculation. However, survival analysis results will reflect whatever bias entered into the chosen threshold crossing time. If poor procedures, language samples, or tests are used, survival analytic results will accurately reflect the poor measurement choices that were made. Unlike traditional methods, however, survival analysis will not compound the problem.

Regression to the Mean

Another statistical problem in longitudinal research is regression to the mean. Scores that are far from the mean tend to move toward the mean in subsequent measurement. Because the combination of factors (known and unknown) that co-occur to contribute to an extreme score is not likely to recombine in subsequent measurement, the score can be expected move toward the mean (Furby, 1973).

Regression to the mean can be confused with therapeutic success and learning. Tomblin, Zhang, and Buckwalter (1997), in a preliminary report of a well-controlled, stratified epidemiological sample of 203 children screened primarily for language delay, found that *all of the improvement* in language scores they measured from kindergarten to second grade could be accounted for by regression to the mean alone.

Participant selection based on extreme scores may also be biased by regression to the mean. A candidate may qualify for inclusion in a study because of an extreme score. Were the measure repeated, the candidate might not qualify because of regression to the mean.

Survival analysis involves a single measure for each participant, so there can be no *statistical* regression to the mean. However, only extraordinary care in the design and analysis of studies can surmount the participant selection problem.

Some Characteristics of Survival Analysis

Survival analytic studies yield individual results. In survival analysis, there is no averaging or summing across individual scores. Thus, results can be applied to individuals in a clinical setting. This contrasts with group analysis—the results of articulation or language tests are typically compared to grouped performances (e.g., means, standard deviations, mental ages, stanine scores, or percentiles). With survival analysis, a report could provide an estimate of the chances of an individual achieving a designated normal range score at some later time. It would then be possible to estimate how rates of improvement and clinical outcomes would most likely change, depending on the type and extent of intervention provided.

Because survival analysis was developed for clinical application, few changes in standard clinical procedures in speech pathology would need to be made to conduct a clinical research study. Clinics may have maintained records suitable for retrospective studies using survival analysis. There are several essential components: The group studied must be a sample of a well-defined population. Some consistent and well-defined outcome measure must have been recorded, at fairly consistent, reasonably closely spaced intervals, until most (but not all) of the sample achieved a specified outcome.

Of course, individuals in a study group should have been treated equivalently during the study. For example, all participants should have received the same therapy. However, if two different therapies were used, and there was a large enough number of outcomes in each therapy group, then it might be possible to divide the participant sample according to therapy type. The two therapies could then be compared for efficacy and efficiency.

To understand how individuals change with or without intervention, longitudinal studies are required. Such studies pose practical problems (e.g., subject retention), but survival analysis is relatively tolerant of the loss of study participants and other limitations of these studies.

Censoring

Censoring occurs when individuals in a study sample are lost to the study or fail to display the target outcome behavior. There are four basic categories of censoring: right, left, interval, and informative. Each category is defined in Table 1. Uncensored outcomes are the basis for probability calculation in survival analysis. *The number of uncensored outcomes is considered to be the size of the study sample, not the number of participants entering a study.*

Although censoring creates serious problems for traditional data analysis, it poses far less difficulty for survival analysis, provided a reasonable contingent of uncensored qualitative changes remain (Allison, 1984). In survival analysis, all right-censored data are utilized. A variety of techniques is also available to estimate gaps in data caused by interval censoring (Klein & Moeschberger, 1997; Pan, 1997). However, informative and left censoring can create insoluble problems (Collett, 1994).

Efficacy and Efficiency Applications

Fiscal restraints in educational and health care settings have had an impact on the delivery of services provided by speech-language pathologists. Included in these trends are changes in case selection and dismissal criteria, increased third-party billing, and pressures to provide better and more objective outcome measures (Taylor, 1992; Trace, 1995; Trulove & Fitch, 1998). Attention will need to be directed toward the ability of speech-language pathologists to formulate more precise and complete prognoses and to cite more specific and objective evidence for efficacy and service efficiency (Pearson, 1995). Although "a prognostic statement is not only required by the ASHA Professional Services Board standards but is required by...fair play and logic,"

Table 1. Definitions for the four categories of censoring and for uncensored outcomes in survival analysis.

Category	Definition
Left censored	The outcome of interest has occurred for a participant before observations for a study have begun.
Uncensored	The outcome is observed during the study window.
Right censored	The outcome of interest has not been observed for a participant after the last observation for a study has been made.
Type I	...because the outcome occurred after the study was over or the outcome never occurred for a participant.
Type II	...because the study was aborted before the outcome of interest occurred for a participant.
Type III	...because the participant was lost to the study before the outcome of interest was observed.
Interval censored	The event of interest for a participant occurred between relatively wide-spaced observations.
Informatively censored	Observation of the event of interest was not possible or the time of observation changed for a reason associated with the nature of the outcome itself.

(Petersen & Marquardt, 1994, p. 303), these statements must be specific so that those who are fiscally responsible can compare alternatives on a dollars-and-cents basis (ASHA, 1992; Pearson, 1995). The American Speech-Language-Hearing Association has established a National Outcomes Measurement System (ASHA, 1999). Survival analysis should be among the methods considered for analysis of the outcome data.

Survival Analysis

The basic approaches to survival analysis are nonparametric, semiparametric, and parametric. Nonparametric and parametric approaches model the time it takes for an outcome or event to occur. Semiparametric approaches, also known as the proportional hazards model or the Cox regression model, sacrifice specific outcome times in order to sharpen tests of differences among groups or treatments.

Table 2 is a summary of the types of survival analysis, including a brief guide for selection of the most appropriate model. Nonparametric survival analytic approaches are appropriate when sample sizes (the number of uncensored outcomes) are small. In survival analysis, “small” is usually interpreted to be in the hundreds or less, although other considerations such as the spacing of outcomes across a study window may also be taken into consideration. Interval censored data, when the interval between measurements is fairly large compared to the duration of the study window, also dictates a nonparametric treatment. Time in some studies is scaled in interval-length units, so nonparametric approaches also have become known as discrete-time methods. If time is measured continuously, parametric or continuous-time approaches are appropriate (Allison, 1984).

To estimate when a given level of performance has been reached, it is often necessary to probe responses at intervals (i.e., interval censoring). Interval censored outcome times are inexact. Because it is impractical to

Table 2. The types of survival analysis with annotations.

Type of survival analysis	Recommended for use when	Introductory reference(s)
I. Nonparametric	1. Sample size (outcomes) is small to large.	
	2. All observations are measured from the same starting point.	
	3. Time (or an equivalent notion) to a change of state outcome is important.	Allison, 1984; Harris & Albert, 1991
A. Life table	Observation windows are of equal size and are equally grouped.	Allison, 1984; Nelson, 1982
B. Kaplan-Meier	Observation windows are of unequal size.	Allison, 1984; Collett, 1994
II. Semiparametric (Cox)	1. Sample size is moderate to large.	
	2. Hazards are proportional.	
	3. Differences between factors or variables are important.	
	4. Time (or an equivalent notion) to an outcome is not important.	Cox & Oakes, 1984; Teachman, 1983
III. Parametric	1. Sample size is large.	
	2. All observations have the same starting point.	
	3. Estimation of the distribution is important.	
	4. Modeling of factors and variables is important.	Blossfeld & Rohwer, 1995; Collett, 1994; Lee, 1992

recruit thousands of participants and continuously monitor each, longitudinal speech-language studies are usually subject to these limitations. Therefore, the appropriate survival analytic methods for application in speech-language pathology research are almost certain to be discrete-time. The following discussion and examples focus on nonparametric approaches. Semiparametric and parametric approaches are only briefly mentioned.

Functional Relationships

In classical survival analysis, the research purpose is to determine how long a patient can be expected to live, given a certain diagnosis or treatment (hence, "survival" analysis). Rephrasing this question to ask the probability of t amount of time passing before death gives the cumulative survival function estimate:

$$\hat{S}(t) = \frac{\text{The number of people surviving to time } t}{\text{The total number of patients}} \tag{1a}$$

To discover how long a patient might be expected to live after contracting some deadly disease, the researcher might examine the death certificates and the medical records from 1,000 individuals. By rank ordering the amount of time it took individuals to die after diagnosis with the disease (death from other causes would constitute right censoring, which reduces the size of the risk set) Equation 1a can be used to provide a sample estimate of the probability of surviving. The time scale for participants may begin upon diagnosis with a disease or condition, upon hospital admission, exposure to a toxin, or initiation of a treatment regimen.

In communicative disorders, the outcome of interest will more likely be a positive one. Normalization is such an outcome. Normalization can be said to occur when an individual who has not previously met a specific set of desired criteria that defines a normal range of speech or language production meets the designated criteria. For communicative disorders, "normalization analysis" would be a more transparent term. Survival time in Equation 1a would be the same as failing to normalize time in Equation 1b below:

$$\hat{S}(t) = \frac{\text{The number of people who did } \textit{not} \text{ normalize by time } t}{\text{The total clinical sample chosen}} \tag{1b}$$

In Equation 1a-b, the value of t is determined for all individuals in a study by the investigator's choice of a zero point in the scaling of time. For communicative disorders the zero point may be the first clinical referral, first assessment, beginning of a school year, and so forth. A developmental study may set the time scale equal for all participants by using chronological age, gestational age, bone age, or achievement of some milestone (e.g., standing unaided, babbling, 50-word vocabulary, school entrance, puberty).

A word of caution is in order concerning the selection of a zero point for the beginning of a study. Suppose a researcher requested that children with specific language impairment (SLI) be referred for a planned study. If testing intervals in the planned study were 4 months apart and the children referred were the same age, give or take 2 months, there is no problem. However, if the children referred varied in age by much more than half the testing interval size, a serious problem called *left truncation* could occur. Suppose further that some children were 3 years old when referred, whereas others were as old as 5. Some children between 3 and 5 years old might have been referred had they been diagnosed at age 3, but were not referred because they had normalized by the time they were noticed by referral sources. When this happens, the sample of children will not include some children who normalized at the younger ages. Therefore, the study will result in a false picture of the age of normalization in SLI children (Klein & Moeschberger, 1997).

The choice of a zero point for the time scale does not preclude employing other temporal measures. They may be introduced as factors or variables. For example, chronological age may be a variable in a study that examines the time it takes for a treatment to produce results.

The probability of normalizing (the failure function) is the complement to Equation 1b:

$$F(t) = 1 - S(t) \tag{2}$$

The cumulative failure function, F(t), is used in medical applications to determine the probability of death at a given elapsed time. In industrial applications, it is used to determine and study product life. With positive outcomes, the same probabilities will correspond to recovery or normalization.

There are six functions of interest in survival analysis: the cumulative survival function, S(t); the survival density function, s(t); the cumulative failure function, F(t); the failure density function, f(t); the cumulative hazard function, H(t); and the hazard density function, h(t). Capital letters are used here to denote cumulative functions; lower-case letters denote density functions. The relationship between the survival and failure functions is a reversal of the time scale.

The hazard density function for a sample is defined as:

$$\hat{h}(t) = \frac{\Sigma d_{ij}}{(\Sigma s_i)\,(w_j)} \tag{3}$$

where $\Sigma d_{i,j}$ = the total number of individuals who normalized (died or experienced an outcome, indicated by the subscript i) in the time period w_j (interval width, where the interval is indicated by the subscript j), Σs_i = the total number of individuals who failed to normalize (i.e., survived).

The hazard density function is the outcome (i.e., normalization, death) *rate* for each time interval. In discrete-time (i.e., nonparametric) analysis, the hazard refers to the probability that an event (i.e., normalization, recovery, death) will occur in a particular time interval to an individual, given that individual is at risk (available) during that time interval. The hazard density function generates actuarial tables that list the probability of death per day (week, hour, etc.).

Referring to Table 2, the following analogy illustrates how time is treated differently in different kinds of survival analysis. Imagine the finish line (outcome) in a cross-country race. In the case of life table analysis, the hazard rate would be based on the number of runners crossing the finish line within some broad, but equal intervals, say every 10 minutes. In Kaplan-Meier (hereafter, KM) analysis, the hazard rate would be based on the time span between each successive runner crossing the line. For each time interval, a single runner crossed the line sometime between the runners who came before and after, but the exact time would not be known. In parametric analysis, the exact time each runner crosses the finish line is known.

The relationship of the hazard function to the survival and failure functions is:

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

The hazard is the number of individuals crossing the threshold (finish line) in each time period, f(*t*), divided by the total number of people in the same time period who could have, but did not yet cross the threshold, S(*t*) (the runners still on the course). For nonparametric analysis, this equation shows that the hazard is the probability that an individual will cross the threshold in a given time interval. In continuous-time (parametric) analysis, the time intervals are infinitesimal. This means that the hazard can be a number larger than 1 with no upper bound, so it is not interpretable as a probability. Instead, continuous-time hazard rates are often thought of as the force, or pressure toward, risk or chance of threshold crossing at an instant. From Equation 4, it can be seen that the six functions are different perspectives on the relationship between time and outcome for individuals at risk of experiencing an event. The six functions can be transformed from one to another.

Life Tables, the Actuarial Approach

The model in Equation 1a-b underlies the life table. The life table is a technique to provide nonparametric estimates of a function. For each data point, a characteristic life table contains the proportions dying (i.e., normalizing or recovering) and continuing (i.e., failing to normalize or recover); the cumulative proportion of participants continuing at the beginning of the interval (continuing function); the conditional prospect (hazard) function, which is the rate of normalization before a specified time interval; and the density function, which is the probability of normalizing within a specified time interval (Allison, 1984; Collett, 1994; Palloni & Sorensen, 1986). The cumulative proportion surviving (failing to normalize, survival function) can easily be calculated, but is not often presented in life tables.

The construction of life tables is straightforward as can be seen using data presented by Kumin, Councill, and Goodman (1994). The speech status of 60 children with Down syndrome (0;9 to 9;0) was examined at 3-month intervals to determine the first time individual consonant phonemes emerged in the speech of each child. The number of children who exhibited emergence for each sound was reported in yearly intervals. Thus, the reported data can be reanalyzed in life tables. Because there are too few outcomes, they cannot be fully modeled.

It might be of interest to know when a phoneme, for example /s/, is likely to emerge in the speech of a child with Down syndrome. To find out, we would subject the data to an actuarial analysis by arranging it as in Table 3. Column 1 numbers the age intervals from youngest to oldest. Column 2 is the interval size in months. Column 3, d_j, shows the number of children whose transcripts showed the outcome, i.e., the emergence of /s/ within the time span shown in Column 2. Column 4, c_j, shows the number of children in each age range who were right censored. According to these data, 13 children never had /s/ emerge in their speech before the study ended, so all are listed in the last time period. No one dropped out of the study. Column 5, n_j, is a list of the total number of children for whom /s/ did not emerge in speech at each time period. (Except for the last row, this number will be the same as Column 5, n_j, minus Column 4, c_j, since right-censored children are no longer in the study). Column 6, n_j', is a list of the number of remaining children at each age period who have not had /s/ emerge in their speech. If we assume that censoring occurs uniformly throughout the interval j′, the average number of individuals at risk during this interval is:

$$n_j' = n_j - \frac{c_j}{2} \tag{5}$$

This is known as the *actuarial assumption*, so 13/2 or 6.5 is entered in the last row of this column.

Column 7, $(n_j' - d_j)/n_j'$, is a list for each age interval of the probability that a child in the study left the interval without having had /s/ emerge. This is calculated as indicated by the column heading. Column 7 may also be

Table 3. Construction of life tables. Data are from Kumin et al. (1994).

Interval number	Interval age range in months	d_j	c_j	n_j	n_j'	$(n_j' - d_j) / n_j'$	S(*t*)	F(*t*)	*SE* S(*t*)
1	0–11	0	0	60	60	1.0	1.0	0.0	0.0
2	12–23	4	0	60	60	.933	1.0	0.0	.036
3	24–35	20	0	56	56	.643	.933	.067	.071
4	36–47	15	0	36	36	.583	.600	.400	.070
5	48–59	4	0	21	21	.810	.350	.650	.066
6	60–71	1	0	17	17	.941	.283	.717	.065
7	72–83	1	0	16	16	.938	.267	.733	.063
8	84–95	1	0	15	15	.933	.250	.750	.062
9	96–107	1	0	14	14	.929	.233	.767	.060
10	108–119	0	13	13	6.5	1.0	.217	.783	.060

Note. The subscript j indicates “within the time interval.” d_j = the number experiencing the event (e.g., dying or /s/ emerging) in the interval. c_j = the number right censored (leaving the interval without experiencing the event). n_j = the total number not experiencing the event in the interval (alive, for whom /s/ did not yet emerge). n_j' = the average number in the interval who remain at risk of experiencing the event. See text concerning the actuarial assumption. $(n_j' - d_j) / n_j'$ = the probability of surviving (not dying, not having /s/ emerge) in the time interval. This corresponds to the survival density function, S(*t*).

interpreted as the outcome probability density function, s(*t*), corresponding to the cumulative outcome probability in Column 8, S(*t*).

If each proportion in Column 7, $(n_j' - d_j)/n_j'$, is multiplied by the probability in Column 8, S(t), (in the same row), the result is the probability that a child will *not* have had /s/ emerge by the next age range. In the first row, the initial probability will always be 1 since, at the beginning, left-censored individuals cannot be included. This allows construction of the cumulative survival function, S(*t*), of Column 8. Because the figures in Column 7, $(n_j' - d_j)/n_j'$, and Column 8, S(*t*), range from 0 to 1, and the question being addressed is the probability at each age range that a child with Down syndrome will have /s/ emerge, we can find this by subtracting each probability S(*t*) from 1, according to Equation 2. The result is F(*t*) (Column 9). Figure 1 is a graph of results from the calculations in Table 3.

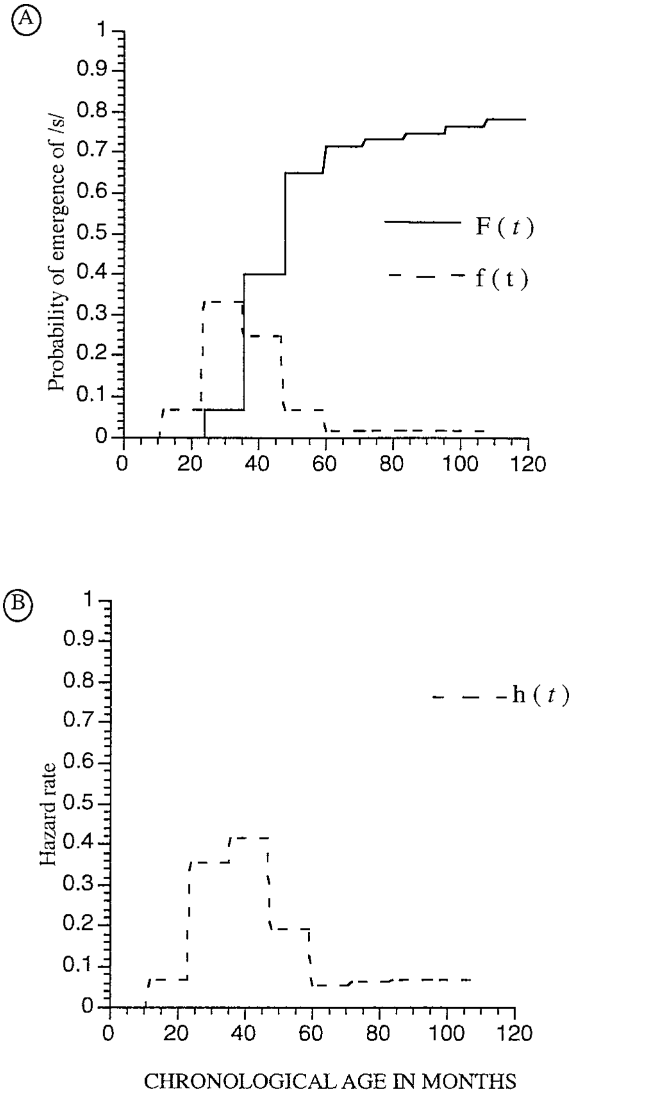
The rightmost column in Table 3, *SE* S(*t*), is the standard error for S(*t*) for each age range. This is calculated as in Equation 6 by taking the square root of the product of the cumulative failure, F(*t*), and survival functions, S(*t*) in Table 3, divided by the sample size, *N* = 47. In survival analysis, sample size is always the number of uncensored outcomes, *not* the number of participants in a study.

$$SE = \sqrt{\frac{S(t)\ F(t)}{N}} \qquad (6)$$

The estimate for the standard error of the survival function is placed in the row above the row in which the calculation was made, since that is the time period when the error occurred. Recall that the calculations for S(*t*) occurred in the row below the one in which calculations were made.

The original question of when /s/ is likely to emerge in the speech of individuals with Down syndrome is answered. Column F(*t*) of Table 3 is a list of the probabilities, accumulating annually up to 10 years of age. For example, the chances are 4 out of 10 [Column F(*t*), Row 4] with standard error of plus or minus about 7 out of 100 [Column *SE* S(*t*)] that /s/ will emerge by 47 months of age (Row 2). The chances of /s/ emerging improve to 6.5 out of 10 by 59 months of age (Row 5), and so on. The probability of /s/ emerging in each year f(*t*), the dashed line in Panel A of Figure 1, can be calculated by subtracting successive entries in the cumulative column F(*t*). Panel B of Figure 1 is the hazard function, h(*t*). This panel is a graphic of the rate of emergence of /s/ in the Down syndrome children in each year of the study. Calculation of h(*t*) was accomplished according to Equation 3. Panel B shows the probability that each Down syndrome child would have /s/ emerge in each yearly interval, provided they had not had /s/ emerge previously. Based on this analysis, we know (a) F(*t*), the overall probability of /s/ emerging in the speech of individual Down syndrome children by the time they reach any age to age 10; (b) f(*t*), the (unconditional) probability for each age to age 10; and (c) h(*t*), the (conditional) probability for an individual child at each year of age who hasn’t already had /s/ emerge. Provided the research sample represents the population seen in a given clinic, these estimates are directly applicable to individual children in that clinic. The reader can repeat the above calculations for the remaining 23 sounds (see Kumin et al., 1994).

Figure 1. Panel A shows the step functions for the cumulative normalization function, F(*t*), and the normalization density function, f(*t*), taken from Table 3, based upon the data reported by Kumin et al. (1994). Panel B shows the hazard density function, h(*t*).



Compare these results to the original report (Kumin et al., 1994). They reported the mean, median, and modal ages at which each of 24 consonant sounds emerged in the speech of their children with Down syndrome. They also reported how many children showed emergence of each sound in each age period.

Kaplan-Meier (KM) Analysis

Actuarial analysis requires that all participants in a study be observed at the same time; otherwise, the times of observation must be grouped. In practice, it may not be possible to observe all participants in a study almost simultaneously. The primary difference between the actuarial approach and the KM approach is that the actuarial approach groups outcomes into equal intervals, whereas the KM approach rank orders outcomes. *In KM analysis, there is a single outcome (or tied outcomes) in each (time) period.* The duration of the period may vary. *In the actuarial approach, the period is fixed and the number of outcomes is variable.* Otherwise, the two techniques are similar.

Kaplan and Meier (1958) provided an estimator for S(t) called the product-limit estimator. This estimator assumes that observed event times are arranged in an increasing order, $t_1 < t_2 <...< t_d$, where d is the total number of events observed. If $N(t_j)$, where j indicates the time interval, are individuals observed not to have normalized (i.e., are at risk) at t_j, d(t_j) normalize (die) at t_j and c(t_j) are censored (do not normalize) in the interval (t_j, t_{j+1}), the KM estimator is defined as:

$$S(t) = \prod_{(j:t_j< t)} \frac{1 - d(t_j)}{N(t_j)} \qquad (7)$$

with the associated estimator of the conditional prospect rate (hazard rate) being the ratio d(t_j) to N(t_j). In KM analysis, intervals are regarded as independent events.

The KM approach is illustrated in the following example (Weiner & Wacker 1982). The study followed 10 normal children and 10 children with severe speech delay over three 6-month intervals. This example is focused on correct articulation of [z] in the word zipper among the children with speech delay. Table 4 is constructed in the same manner as actuarial Table 3. The conventions and notations used are identical to those employed in the actuarial table.

Column 1, "Case," is the number assigned to participants ranked from the youngest to the oldest age at which the [z] was correctly articulated or at which the subject was censored. Column 2 is a list of the age in months for each child at the session during which [z] was first correctly articulated. Notice that Cases 6 and 7 occurred at the same age.

Column 3, c_j, is the censoring status for each participant at the outcome age in Column 3 (age). The "LC" for Child 1 at 51 months of age indicates the child was left censored. He produced [z] correctly the first time he was tested. Consequently, all we know is that sometime from birth to age 51 months, this child acquired the ability to articulate correctly the [z] in zipper. Because the size of this uncertainty is so great, this child is omitted from the analysis. Left censoring such as this can be a serious threat to the validity of a study. The next 2 subjects were right censored at 58 and 59 months of age, respectively. Child 2 was right censored after being tested only twice. (He moved away from the area.) This is Type-III right censoring (see Table 1). Child 3 was tested three

Table 4. Construction of a Kaplan-Meier table. Data were taken from Weiner and Wacker (1982). The outcome is correct pronunciation of the [z] phone in "zipper."

Case	Age (months)	c_j	d_j	n_j	$(n_j - d_j) / n_j$	S(*t*)	F(*t*)	*SE* S(*t*)
1	51	LC	—	—	—	—	—	—
2	58	C	0	9	1	1	.000	.000
3	59	C	0	8	1	1	.000	.000
4	63	U	1	7	.857	.857	.143	.132
5	65	C	0	6	1	.857	.143	.132
6	68	U	1	5	.800	.514	.486	.204
7	68	U	1	4	.750	.514	.486	.204
8	76	C	0	3	1	.514	.486	.204
9	78	U	1	2	.500	.257	.743	.208
10	82	U	1	1	0	0.0	1.000	0.0

times; then the study ended. He still had not articulated the [z] correctly. This is right censoring, Type I.

It is not necessary to identify the type of right censoring in KM analysis. All right censoring is usually just labeled "censored" or "c." The fourth child was uncensored, "u." He first uttered [z] correctly at 63 months, at his second testing. All outcomes were interval censored because the children were tested at 6-month intervals. Interval censoring can cause serious methodological problems. The oldest age in each interval could be arbitrarily taken as the age at which the children acquired the [z] sound. This strategy is often used in KM analysis to cope with interval censoring. It provides the most conservative, if not the most accurate, estimates. An alternative is to use the midpoint of each interval so that the standard error is bidirectional, not unidirectional as it would be if interval endpoints were used.

Column 4, d_j, corresponds to the same column in actuarial Table 3. In KM analysis, this is restricted to 0 or 1. In actuarial analysis, any number may be entered into the rows. The number in this column is the number of children correctly producing [z]. Because each child is listed separately, the number is a 1 if the child correctly produced the sound and a 0 if the child was right censored.

Column 5, n_j, is the number of children who remain in the study (were not right censored and remain at risk) and who have not yet produced [z] correctly. With only a single individual in each time interval, it is no longer necessary to distinguish n_j from n_j'. Columns 6 through 9 are calculated in the same manner as in the actuarial example. An exception occurs when outcome times are tied, as happened for Participants 6 and 7. When this occurs, S(*t*) must also be the same, although the proportion $(n_j - d_j)/n_j$ will be calculated for each case. The highest ranking (last calculated, longest time-span) figure is applied retroactively to the preceding case (time) as can be seen for these rows in the rightmost three columns of Table 4.

Column 8, F(*t*), is a list of the cumulative probabilities that a child in this study articulated the [z] in zipper correctly by the age in Column 2. Column 6, $(n_j - d_j)/n_j$, is the corresponding density function, s(*t*). These results are presented in Figure 2.

Panel A of Figure 2 is a representation of the survival density function, s(*t*), from Column 6 of Table 4, $(n_j - d_j)/n_j$, and the cumulative survival function, S(*t*), from Column 7. Panel B is a display of the corresponding failure density function, f(*t*), which is not presented in Table 4, and, from Column 8 of Table 4, the cumulative failure function, F(*t*). Because the outcome concerned is of a positive nature, the failure functions in Panel B represent the probabilities of the first *correct* use of the /z/ in zipper. The survival functions in Panel A represent the probabilities that individuals will not yet have uttered /z/ correctly.

The dotted line in Panels A and B that occurs between the ages of 68 and 76 months is provided to show the two outcomes between these ages for Cases 6 and 7 from Table 4. The hazard functions are presented in Panel C. The hazard density function is calculated from the survival and failure functions by using Equation 4.

The hazard density function, h(*t*), represents the *rate* (per month) at which the first correct use of the /z/ in zipper was recorded in each time interval. This function is known as the age-specific failure rate, or the *force* toward the outcome. The solid line represents the (integrated) cumulative hazard rate, H(*t*). The cumulative hazard rate is related to the cumulative survival function, S(*t*):

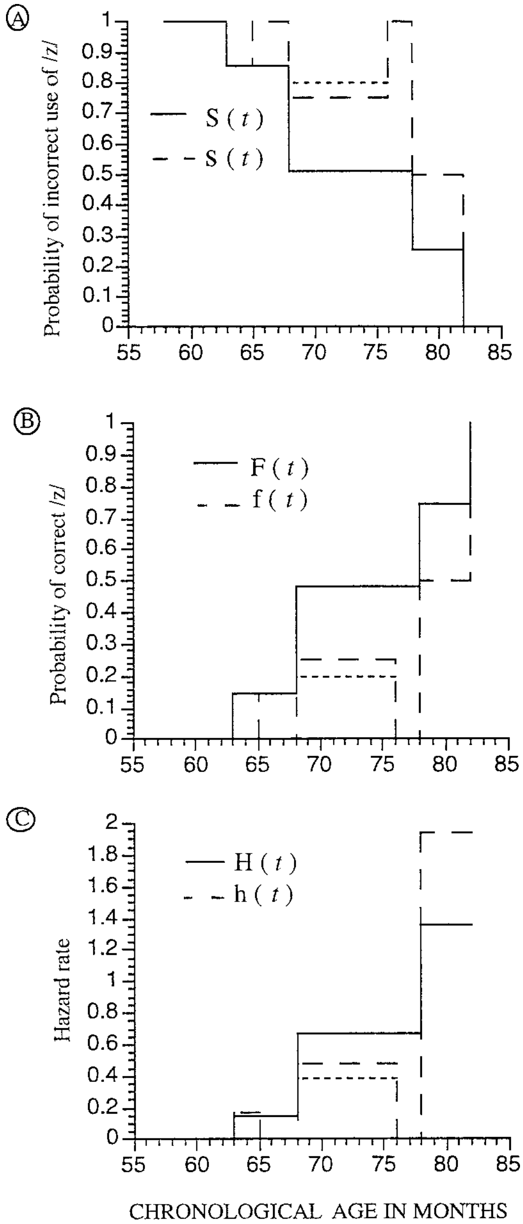
$$H(t) = -\log_e S(t) \tag{8}$$

By substitution, it is also true that:

$$f(t) = h(t) \exp[-H(t)] \tag{9}$$

In Panel C, unlike the survival and failure functions, which can only assume values from 0 to 1, both hazard functions can take any positive value. This is because they are rates, not probabilities. It is also possible, as can be

Figure 2. Panel A shows the survival functions, S(*t*) and s(*t*), which indicate the probabilities an individual had not yet achieved correct articulation of /z/. Panel B shows the normalization (failure) functions, F(*t*) and f(*t*), which indicate the probability of an individual achieving correct articulation of /z/. Panel C shows the hazard functions, H(*t*) and h(*t*), which show the aggregate and age-specific rates of first correct articulation of /z/. Data are from Table 4 based on a report by Weiner and Wacker (1982).



seen for the last age interval (from 78 to 82 months), that the cumulative hazard rate, H(*t*), can be smaller than the density hazard rate, h(*t*). This is because accumulation, or integration, takes into account previous smaller rate values. In panel C, H(*t*) is smaller (slower) than h(*t*) at the last age interval because the previous intervals (cf. from 76–78 months) were zero. The cumulative hazard rate, H(*t*), is interpreted as the overall rate of the outcome measure, accumulated up to and including each successive time interval.

Life table estimates approach KM estimates whenever time intervals are small and/or the number of events is large. The KM estimates of S(*t*) and the conditional prospect (hazard) are influenced by biases for small samples, but corrections are available (Aalen, 1978; Nelson, 1982). These estimates have been shown to have good large sample properties (Kaplan & Meier, 1958).

Some Methods for Group Comparison

In survival analysis, groups can be contrasted easily. This feature is advantageous in efficacy studies, clinical trials, and basic research. This section is a presentation of some common methods for comparing groups.

The example selected was designed to evaluate the efficacy of a therapy called "Metaphon" (Dean, Howell, Waters, & Reid, 1995). The units of phonological analysis in Metaphon are natural phonological processes (Stampe, 1969, 1972). Thirteen preschool children with speech delay participated for an average of 17 half-hour sessions held once a week. Change was assessed every third session. All of the children completed the study, so only Type-I right censoring occurred. Before the study began, each child was assessed, and baseline measures for natural phonological processes were established. For each child, three processes that were operating at a 100% level were chosen. One of these processes served as a control throughout the entire study. The other two processes were targets for Metaphon treatment from the same experienced clinician. The two treatment processes were handled sequentially. A phonological process was selected as the first target for therapy, which was conducted until the process was suppressed at least 50% of the time. After reaching this criterion with the first process, treatment began on the second process, which was treated until it also reached this criterion or until the study ended. No therapy was targeted for the control process, although its course was assessed as in treatment.

The authors only reported the processes that were chosen for 9 of the 13 children for whom therapy success was found, so no reanalysis of results using processes as a covariate are possible. However, it is apparent from the reported processes that velar fronting was over-represented in the first phase of treatment, and gliding of liquids was over-represented as a control process. Velar fronting was never a control process. Liquid gliding was never treated during the first phase. There is a better balance among process types for the second phase of therapy and the control processes, so the example will use only the results from the second phase of therapy.

Table 5. A Kaplan-Meier table for suppression of phonological processes before and after Metaphon therapy. Data are from Dean, Howell, Waters, and Reid (1995).

	Control processes					Treatment processes			
(Case) Time	n_i	d_i	c_i	S(*t*)	(Case) Time	n_i	d_i	c_i	S(*t*)
(13) 4	12	1	0	.923	(6) 5	12	1	0	.923
(6) 5	11	1	0	.846	(4) 6	11	1	0	.846
(9) 6	10	1	0	.769	(1) 7	7	1	0	.538
(2) 7	9	1	0	.692	(2) 7	7	1	0	.538
(10) 8	8	1	0	.615	(8) 7	7	1	0	.538
(1) 9	3.5	1	0	.538	(9) 7	7	1	0	.538
(3) 10	3.5	0	1	.538	(5) 9	3	1	0	.462
(4) 10	3.5	0	1	.538	(3) 10	3	0	1	.462
(5) 10	3.5	0	1	.538	(7) 10	3	0	1	.462
(7) 10	3.5	0	1	.538	(10) 10	3	0	1	.462
(8) 10	3.5	0	1	.538	(11) 10	3	0	1	.462
(11) 10	3.5	0	1	.538	(12) 10	3	0	1	.462
(12) 10	3.5	0	1	.538	(13) 10	3	0	1	.462

Table 5 is a KM analysis of the therapy outcomes and the control processes. The level of suppression of the processes employed as a cutoff was 85%. In Column 1, the participant number corresponding to Dean et al. (1995) is shown in parentheses. The session number is shown to the right. There were no outcomes for either group before Session 4, so earlier sessions were omitted.

The notations in this example correspond to those previously employed. Columns 2 and 8, n_i, are the number of children remaining at risk, that is, who have not yet normalized for the control process and the treatment process, respectively. Columns 3 and 9, d_i, are the number of children who suppressed the phonological process (normalized) for control and treatment processes respectively. Columns 4 and 10, c_i, are lists of the censoring status for both groups. Columns 5 and 11, S(t), are the cumulative survival functions for each group. These are calculated as previously described for KM analysis. Figure 3 is a graphic of these survival functions. Because natural process suppression is a negative outcome, as death is for traditional survival analysis, there is no need to invert graphs and the semantics of labels to make them more intuitive.

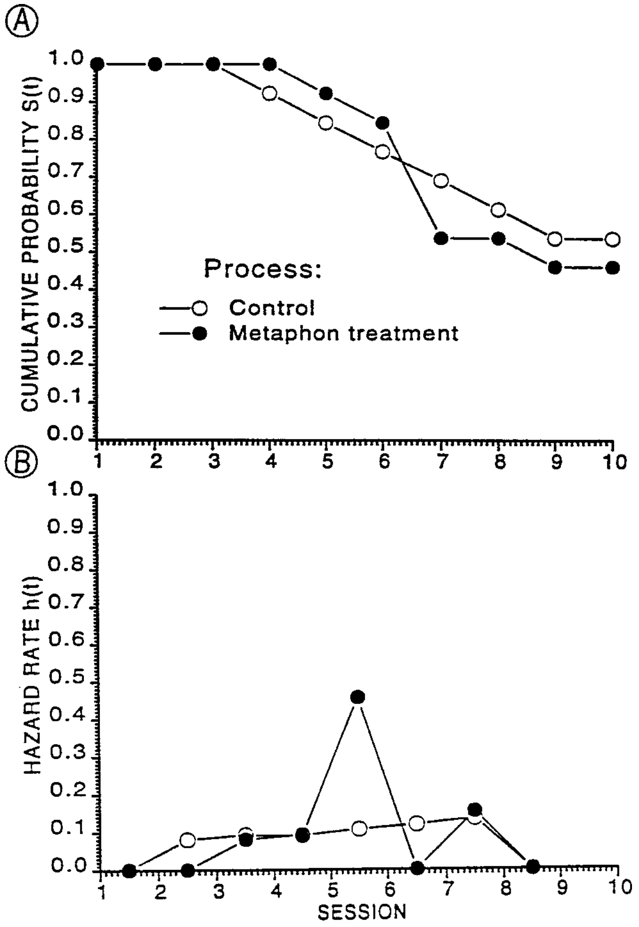
The Relative Risk at Specific Intervals

The relative risk (of normalizing) at each session (time) is:

$$RR_i = \frac{F(t)_{it}}{F(t)_{ic}} \qquad (10)$$

where i indicates individuals, c indicates controls, and t indicates treatment. The relative prospect (risk) of normalizing (see Table 5, 8th week) would be:

Figure 3. Panel A shows the cumulative survival function, S(*t*), for speech delayed children. Outcomes were at the 85% level for control and Metaphon treatment processes. Panel B shows the hazard density function, h(*t*), corresponding to Panel A, from data reported by Dean et al. (1995), as presented in Table 5.



$$RR_8 = \frac{1 - .538}{1 - .615} = \frac{.462}{.385} = 1.47 \qquad (11)$$

Recalling the calculation for the standard error, application to Session 8 (time) from Table 5 gives:

$$z = \frac{.538 - .615}{\sqrt{(.138)^2 + (.135)^2}} = \frac{-.078}{\sqrt{.037269}} = -.404 \qquad (12)$$

With $z = -.404$, assuming a normal distribution and referring to the appropriate tables, the p value is approximately .34, so the difference at this time would not be considered significant. Although individuals in the treatment group were about 1.5 times more likely to normalize, this increased likelihood was not beyond chance.

The Mantel-Cox Logrank Test

The previous test applied only to a single session or time. The Mantel-Cox Logrank Test, most often referred to as the Logrank Test (even though it does not involve either ranks or logarithms), makes use of all of the available data. This test was derived from the Mantel-Haenszel chi-square test and is so distributed. The logrank test is the test of choice in survival analysis. Being commonly applied, it has spawned variations for special problems.

The logrank is a test of the hypothesis of no differences among groups (or treatments) across time (sessions). Based on the same logic and distributed as the chi-square test, it compares the number of events observed with the number expected by chance were there to be no difference among groups.

Table 6 is a reorganization of the data from Table 4 as taken from Dean et al. (1995). Column 1 is a list of the session times. Columns 2–4 are lists of the corresponding numbers of participant processes in both groups and the total number who have not met the outcome criterion and so are still at risk (of normalizing). Columns 5–7 are corresponding lists of the numbers of participant processes that were suppressed (normalized). The last two columns were calculated to show the number of participant processes that would be expected for both the treatment and the control groups were these two groups to have no outcome (suppression, normalization) differences. The expected frequencies are calculated as the chi-square statistic, specifically:

$$E_i = d_i \left(\frac{n_{ix}}{n_{ix} + n_{iy}} \right) \qquad (13)$$

where x and y are factors.

The Mantel-Cox chi-square statistic is:

$$\chi^2 = \frac{(O_t - E_t)^2}{E_t} + \frac{(O_c - E_c)^2}{E_c} \qquad (14)$$

The totals figures from the bottom of Table 6 can now be inserted:

$$\chi^2 = \frac{(7 - 6.5986)^2}{6.5986} + \frac{(6 - 6.4014)^2}{6.4014} = .0496 \qquad (15)$$

By the cumulative chi-square distribution for one degree of freedom, there is a probability of less than .83 that the treatment outcome differed from the control outcome. From these results, we fail to reject the null hypothesis. We cannot say that Metaphon therapy differs from no therapy in terms of how long it takes children with speech delay to suppress natural phonological processes.

The Overall Relative Risk

The relative risk for all intervals is:

Table 6. The Kaplan-Meier outcome table reorganized to facilitate calculation of statistical contrasts between groups. See text for details.

Session	At Risk			Suppressed			Expected	
	Treatment	Control	Σ	Treatment	Control	Σ	Treatment	Control
Time	n_{it}	n_{ic}	n_i	d_{it}	d_{ic}	d_i	E_{it}	E_{ic}
1	13	13	26	0	0	0	0	0
2	13	13	26	0	0	0	0	0
3	13	13	26	0	0	0	0	0
4	13	12	25	0	1	1	.5200	.4800
5	12	11	23	1	1	2	1.0435	.9565
6	11	10	21	1	1	2	1.0476	.9524
7	7	9	16	4	1	5	2.1875	2.8125
8	7	8	15	0	1	1	.4667	.5333
9	7	3.5	10.5	1	1	2	1.3333	.6667
10	3	3.5	6.5	0	0	0	0	0
			Σ:	$O_t = 7$	$O_c = 6$		$E_t = 6.5986$	$E_c = 6.4014$

$$\mathrm{RR} = \frac{\mathrm{O_t/E_t}}{\mathrm{O_c/E_c}} \tag{16}$$

Which, applied to the Metaphon data from Table 6 gives:

$$\mathrm{RR} = \frac{7/6.19}{6/6.81} = \frac{1.1309}{.8811} = 1.2835 \tag{17}$$

Because it will not result in a significant chi-square under one degree of freedom with an alpha of .05, a risk ratio under 2 is not usually considered significant (cf. Blossfeld & Rohwer, 1995; Norman & Streiner, 1994). When a chi-square is not found to be significant in the logrank test, the overall risk ratio is not normally calculated. The overall risk ratio of 1.28 suggests that Metaphon treatment can be expected to benefit about 1 out of 5 children with speech delay. This finding may be due to chance alone.

Other Tests

Although it is the most frequently used and recommended test statistic for use in both parametric and nonparametric survival analysis, the logrank test has limitations. The logrank test should only be applied when the hazard rates for the factors being compared are proportional.

One way of determining whether hazard rates are proportional is to examine the survival plots S(t). If the two groups of survival data do not cross each other, the hazards are proportional. If they do cross, the assumption of proportional hazards is not warranted. Referring back to Figure 3 for Metaphon treatment, the S(t) plots do cross, so the logrank statistic calculated assuming proportional hazards, should *not* have been used.

The appropriate statistic, which does not assume proportional hazards, is the Wilcoxon test. Collett (1994) provided an exposition of the Wilcoxon test and the Gehan, or generalized Wilcoxon test, for use when more than two groups are being compared. Dawson-Saunders and Trapp (1994) suggest a clever method to calculate the Wilcoxon test for two independent samples. In this approach, survival times for each group are ranked from shortest to longest, then the t test is performed on the ranks of survival times. The independent-groups t test is not appropriate because survival times themselves are usually not normally distributed. Most survival times are extremely positively skewed. The two-sample t test for the point biserial correlation coefficient would be more appropriate (cf. Marascuilo & Serlin, 1988, pp. 424–427).

However, the Wilcoxon test is not appropriate if there are censored observations. When there are censored observations, the Breslow-Gehan-Wilcoxon test or the Petro-Petro-Wilcoxon test would be appropriate. The Breslow-Gehan-Wilcoxon test gives greater weights to times with more observations in the risk set. This weighting renders it less sensitive to late events when few participants remain in the study as compared to the Wilcoxon or the logrank tests. Weighting by the number of observations provides an advantage when the data set has tied observations.

Should there be a marked difference in the amount or pattern of censoring among groups, the Breslow-Gehan-Wilcoxon test may lead to anomalous results. The Petro-Petro-Wilcoxon test derives weights from an estimate of the survival function, and the Harrington-Fleming family of tests allows the researcher to control the weights for specific problems or applications. These tests are also available in trend versions. References are Cox and Oakes (1984), Kalbfleisch and Prentice (1980), Lawless (1982), and Nelson (1982). All the appropriate tests were applied to the Metaphon problem. None provided significant results.

As with post hoc multiple comparison procedures in the analysis of variance, the most appropriate comparison procedure depends on the nature of the observations. Consequently, there is no single method used by investigators. In most circumstances, all of the methods of comparison mentioned give similar results.

Adjusting for Covariates

Because the tests discussed are based on the chi-square test, the problem of covariation is readily handled. For example, to discover whether specific types of natural phonological processes had differential effects on survival times in the Metaphon study, the researcher could divide the groups into strata according to the processes of interest, then expand the chi-square table accordingly. A disadvantage is that the sample size for each cell drops accordingly. In the Metaphon study, this is not possible because the number of outcomes is already very small. A second disadvantage is that there is no way to estimate the magnitude of the effect. A third disadvantage is that to form strata, continuous variables may have to be divided into discrete factors. In doing so, power and sensitivity are lost.

Some equivalent to an analysis of covariance for survival data is needed so that continuous data can be treated as continuous and can handle any number of covariates, and also so an estimate of the magnitude of differences can be determined. This is possible, if the hazard rates are proportional, through use of the Cox proportional hazards model.

The (Semiparametric) Cox Proportional Hazards Model

A basic model employed in the analysis of the survival data is the proportional hazard model proposed by

Cox (1972). Although this model assumes no particular probability distribution, it does assume proportional hazards between groups. The hazard of normalization, at any given time for an individual in one group, is proportional to the hazard at that time for a similar individual in the other group. The Cox model is considered semiparametric. Its use follows the principles for parametric model determination. Introductions to the Cox model are in Christensen (1987) and Harris and Albert (1991).

The key to understanding the Cox model is the realization that once the hazards h(t) are found to be proportional, their ratio can be considered a constant. For any time t:

$$\frac{h_i(t)}{h_j(t)} = c \tag{18}$$

where i and j are individuals.

The constant c depends upon the explanatory variables, but not on time, which is partialled out. This makes the Cox model ideal when the focus of research is on the relationship between explanatory variables, but not applicable when the focus is on the dependence of the hazard on time. The Cox model is appropriate for most efficacy study comparisons, but inappropriate for many questions about development. For example, if one suspects that the early development of /s/ with distortions disposes an individual to retain residual /s/ distortions, the Cox model would not be appropriate. In the Cox model, time retains an ordinal quality but loses ratio scaling.

Parametric Models

If the number of outcomes in a study is large enough, it is desirable to model survival data parametrically. A first step is to determine the distribution, then to model factors and variates. Not only are the parameters of intrinsic interest, but precision can be increased by using parametric models (Allison, 1984; Blossfeld, Hammerle, & Mayer, 1986; Carroll, 1982; Kalbfleisch et al., 1980; Miller, 1981; Teachman, 1983). All recent releases of major statistical packages contain routines for survival analysis (e.g., Minitab, 1997; SAS/STAT, 1997; BMDP, 1997a, 1997b).

Clinical Judgment or Probability Estimation?

At present, outcome prediction usually relies on a clinician's intuitions and experience, informed by objective test results, sometimes aided by diagnostic therapy (Winitz, 1984). Suppose that, after appropriate assessment procedures, the best judgment of an experienced speech clinician was in disagreement with results from a sound survival analytic study. What course of action should be followed? One based upon clinical experience or one based upon actuarial research?

Because of the huge number of variables that would be required to construct faithful actuarial predictions, expert clinical judgment may be more trustworthy. However, clinicians may disagree. Meehl (1958) concluded that, "Both theoretical and empirical considerations suggest that we would be well advised to concentrate effort on improving our actuarial techniques rather than on the calibration of each clinician for each of a large number of different prediction problems" (p. 505).

In a review of 65 articles contrasting the judgments of physicians with analytical and actuarial approaches to the same prediction problems, clinical judgment was found to be less accurate than alternative actuarial approaches (Dawson, 1993). Survival analysis can summarize, contrast, and present clinical outcome experience in explicit terms not previously available to speech and language clinicians.

Acknowledgments

Preparation of this article was, in part, supported by Research Grant RO1 DC00496 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Lawrence D. Shriberg, principal investigator. This article is adapted from part of a PhD dissertation at the University of Wisconsin–Madison under the supervision of Lawrence D. Shriberg. This article also benefited from the advice of Robin Chapman, Raymond D. Kent, Colleen Moore, Dolores (Dee) Vetter, Diane Austin, Peter Flipsen Jr., Joan Kwiatkowski, Hye-Kyeung Seung, David Wilson, and three anonymous reviewers.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics, 6*, 701–726.

Allison, P. (1984). *Event history analysis: Regression for longitudinal event data.* Newbury Park, CA: Sage.

American Speech-Language-Hearing Association. (1992). Prescription. *Asha, 34* (Suppl. 9), 9.

American Speech-Language-Hearing Association. (1994). Code of Ethics. *Asha, 36* (Suppl. 13), 1–2.

American Speech-Language-Hearing Association. (1999, January). *National Center on Treatment Effectiveness in Communication Disorders* [Information posted on the World Wide Web]. Rockville, MD: Author. Retrieved March 3, 1999 from the World Wide Web: http://www.asha.org/NCTECD/treatment_outcomes.htm

Blossfeld, H.-P., & Rohwer, G. (1995). *Techniques of event history modeling*, Mahwah, NJ: Lawrence Erlbaum.

Blossfeld, H.-P., Hammerle, A., & Mayer, K. U. (1986).

Ereignisanalyse: Statistische theorie und anwendung in den wirtschatts und sozialwissenschaften. Frankfurt: Campus Verlag.

BMDP (Version 7) [Computer software]. (1997a). Chicago: SPSS, Inc.

BMDP (Version 8.0 for Windows) [Computer software]. (1997b). Chicago: SPSS, Inc.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in contemporary education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.

Carroll, G. R. (1982). *Dynamic analysis of discrete dependent variables: A didactic essay* (vol. 8). Mannheim: Zuma-Bericht.

Christensen, E. (1987). Multivariate survival analysis using Cox's regression model. *Hepatology, 7*, 1346–1358.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Collett, D. (1994). *Modeling survival data in medical research*. London: Chapman & Hall.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, 74*(B), 187–220.

Cox, D., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman & Hall.

Dawson, N. V. (1993). Physician judgment in clinical settings: Methodological influences and cognitive performance. *Clinical Chemistry, 39*, 1468–1478.

Dawson-Saunders, B., & Trapp, R. G. (1994). *Basic & clinical biostatistics* (2nd ed.). Norwalk, CT: Appleton & Lange.

Dean, E. C., Howell, J., Waters, D., & Reid, J. (1995). Metaphon: A metalinguistic approach to the treatment of phonological disorder in children. *Clinical Linguistics and Phonetics, 9*, 1–20.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8*, 172–179.

Gruber, F. A. (1997). *Developmental phonological disorder: A survival analysis*. Unpublished doctoral dissertation, University of Wisconsin–Madison.

Harris, E. K., & Albert, A. (1991). *Survivorship analysis for clinical studies*. New York: Marcel Dekker.

Hayes, W. L. (1973). *Statistics for the social sciences*. New York: Holt, Rinehart & Winston.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: John Wiley.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brook/Cole.

Klein, J. P., & Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. New York: Springer.

Kumin, L., Council, I., & Goodman, M. (1994). A longitudinal study of the emergence of phonemes in children with Down syndrome. *Journal of Communicative Disorders, 27*, 293–303.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. New York: Wiley.

Leonard, L. L., & Orchard, D. (1996). The problem of generalizing to a language population: A "random" controversy. *Journal of Speech and Hearing Research, 39*, 406–413.

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks-Cole.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York: W. H. Freeman.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

Meehl, P. (1958). When shall we use our heads instead of the formula? In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science: Vol. II Concepts, theories, and the mind-body problem* (pp. 498–506). Minneapolis, MN: University of Minnesota.

Miller, R. G., Jr. (1981). *Survival analysis*. New York: Wiley.

Minitab (Version 12 for Windows '95 and Windows NT) [Computer software]. (1997). State College, PA: Minitab, Inc.

Nelson, W. (1982). *Applied life data analysis*. New York: John Wiley.

Norman, G. R., & Streiner, D. L. (1994). *Biostatistics: The bare essentials*. St. Louis, MO: Mosby.

Nunnally, J. E. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Palloni, A., & Sorensen, A. (1986). *Methods for the analysis of event history data* (CDE working paper 86-36). Center for Demography and Ecology, University of Wisconsin–Madison.

Pan, W. (1997). *Non- and semi-parametric survival analysis with left truncated and interval censored data*. Unpublished doctoral dissertation, University of Wisconsin–Madison.

Pearson, V. A. H. (1995). Speech and language therapy: Is it effective? *Public Health, 109*, 143–153.

Petersen, H. A., & Marquardt, T. P. (1994). *Appraisal and diagnosis of speech and language disorders*. Englewood Cliffs, NJ: Prentice-Hall.

SAS/STAT (Version 7) [Computer software]. (1997). Cary, NC: SAS Institute, Inc.

Stampe, D. (1969). The acquisition of phonemic representation. *Proceedings of the fifth regional meeting of the Chicago Linguistic Society* (pp. 433–444). Chicago, IL: Chicago Linguistic Society.

Stampe, D. (1972). *A dissertation on natural phonology*. Unpublished doctoral dissertation, University of Chicago.

Taylor, J. S. (1992). *Speech-language pathology services in the schools*. Boston: Allyn & Bacon.

Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards models. *Social Science Research, 12*, 263–301.

Tomblin, B. J., Zhang, X., & Buckwalter, P. (1997, May). *A statistical expectation for the recovery rate of SLI from kindergarten to second grade*. Paper presented at the 18th Annual Symposium on Research in Child Language Disorders, Madison, WI.

Trace, R. (1995). Outcome measures lead to increased accountability by practitioners. *Advance. 5*(32), 6–7.

Trulove, B. B., & Fitch, J. L. (1998). Accountability measures employed by speech-language pathologists in private practice. *American Journal of Speech-Language Pathology, 7*, 75–80.

Weiner, F. F., & Wacker, R. (1982). The development of phonology in unintelligible speakers. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 8). New York: Academic Press.

Willett, J. B., & Singer, J. D. (1989). Two types of question about time: Methodological issues in the analysis of teacher career path data. *International Journal of Educational Research, 13*, 421–437.

Willett, J. B., & Singer, J. D. (1991). How long did it take? Using survival analysis in educational and psychological research. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 310–328). Washington, DC: American Psychological Association.

Winitz, H. (Ed.). (1984). *Treating articulation disorders: For clinicians by clinicians*. Baltimore, MD: University Park Press.

Received March 10, 1998

Accepted July 23, 1998

Contact author: Frederic A. Gruber, PhD, Lamar University, Department of Communication Disorders, PO Box 10076, Beaumont, TX 77710. Email: gruberfa@hal.lamar.edu