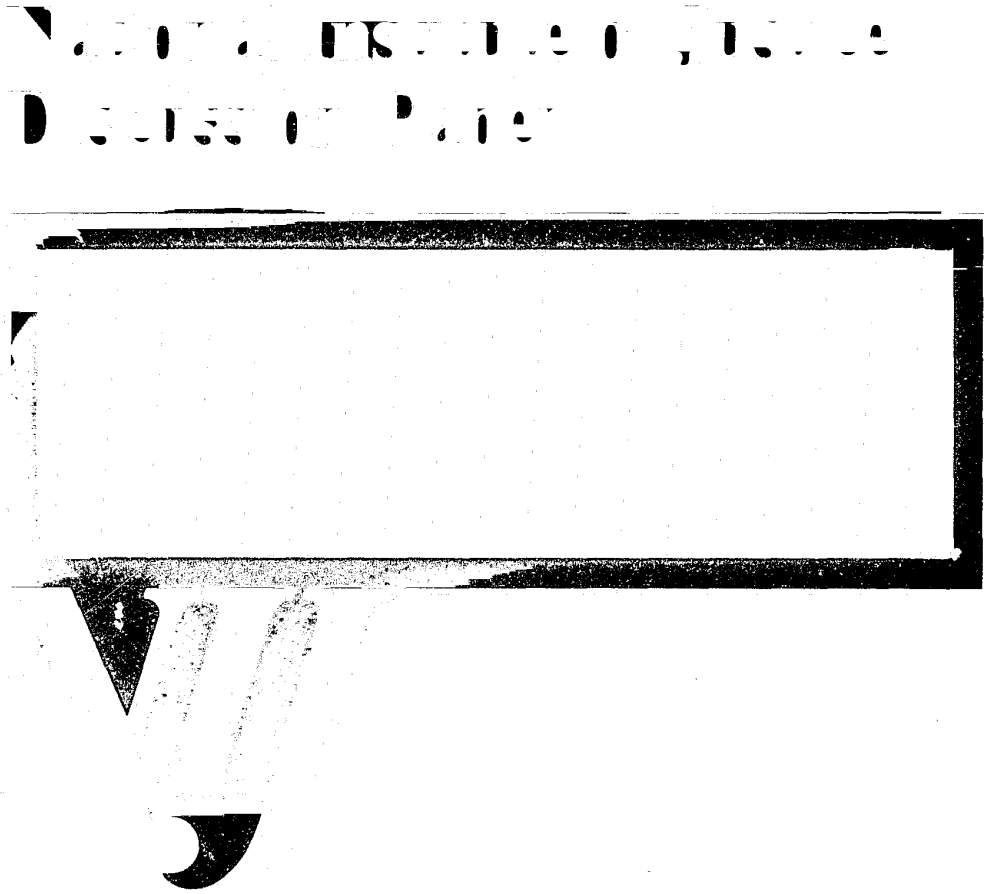

Probability Models of Recidivism: An Exploration

Richard L. Linster
National Institute of Justice

E. Britt Patterson*
School of Criminology
Florida State University

Discussion Paper 3-87

*Supported in part by NIJ Grant #84-IJ-CX-K046 to University of Maryland.

Opinions expressed in this discussion paper are those of the authors and do not necessarily reflect the views of the U.S. Department of Justice.

CONTENTS

I. INTRODUCTION 2

II. A BIT ON MODELING MATHEMATICS 7
A. On Model Forms 7
1. The Logit Model 7
2. Hazard Models 8
B. On Likelihood Estimation 11
Note on the Gamma Hazard Model 14

III. THE ANALYTIC APPROACH 17
A. The Overall Scheme 17
B. Model Construction 20
C. On Model Validation 26

IV. THE NORTH CAROLINA DATA 29
A. The Data Base 29
B. Logit Analysis 32
C. Hazard Models -- The Proportional Hazard Solution . 42
D. The "Full" Hazard Model Solution 46
E. Speculations on Policy Use of Hazard Models 54

V. CALIFORNIA YOUTH AUTHORITY DATA 63
A. The Data Base 63
B. Logit Analysis 69
C. Hazard Models 73
1. The Proportional Hazard Model 73
2. The "Full" Hazard Model 77

VI. SOME CONCLUDING REMARKS 91

References 96

I. INTRODUCTION.

This paper is about predicting recidivism. More precisely, it is about the use of actuarial models that seek to associate a failure probability with a set of selected attributes of an offender and his history.[1]

Perhaps the simplest and most frequently encountered use of actuarial methods occurs in evaluative studies in which statistical "controls" are introduced as a replacement for or supplement to a true experimental design. In a certain sense such studies may have more of a "historical" than an explicitly predictive purpose. They are concerned with achieving a careful and succinct summary understanding of what was observed to happen at a particular time and under a particular set of conditions. In this, of course, they demonstrate that spirit of objectivity and caution in generalization that characterizes all scientific research. However, they differ from research reports in the physical and biological sciences, say, in two important respects.

First, criminology is still groping for a well tested set of theories that would reduce the great uncertainty that exists in determining what exogeneous conditions are important to specify in order to draw reliable inferences about intervention effects from actuarial models. In most cases a randomized assignment of subjects into treatment and control groups would serve to obviate that problem. But true experiments are notoriously difficult to implement and maintain within the operating criminal justice system. Which is, of course, why researchers must so often turn to statistical models.

A second difference lies in the expectation that a study will be replicated. Repeated tests of findings are the rule in the laboratory sciences. For many reasons, however, careful replications of social science studies in general and recidivism studies in particular are quite rare. As a consequence, statistical methods for inferring something about the generalizability of results from a single study have received a great deal of attention and some of these methods have found their way into the standard operating procedures of the recidivism modeling process. Thus, for example, something about generalizibility is being conveyed in the reporting of the t-statistics associated with a model's parameters and

---

[1]This paper will not attempt to give an overview of the very large and diverse literature dealing with this topic. Fortunately, excellent critical reviews are contained in a number of works published quite recently. (Farrington and Tarling, 1985; Gottfredson and Gottfredson, 1986; Gottfredson and Tonry, 1987.)

characterizing as "not significant" those variables whose coefficients have t values below some specified number. Justification for any confidence placed in the predictive power of such models -- that is, analytic support for the assertion that the findings would be similar if the study were in fact replicated -- rests entirely on the assumptions and the logic of statistical distribution theory.

More explicitly predictive in their stated purpose are those studies aimed at improving some specific class of criminal justice decisions. A simple example for recidivism prediction is the study in which individuals are assigned a risk score based on information pertinent to the prognosis of success and available at the time a decision such as parole release is to be made. Individuals whose scores fall below a pre-determined cutting point are the predicted successes; those with higher risk scores the predicted failures. The evaluation of such a prediction instrument consists in observing which subjects actually did fail during some follow-up period and calculating rates of true and false predictions.[2]

There would seem to be two possible justifications for this latter approach. Conceivably, the recidivism process is fundamentally deterministic. An individual is either a recidivating type or he is not. The model tries to distinguish the sheep from the goats and the evaluation tests how well it succeeds. Researchers may not subscribe literally to such an interpretation of what they are striving to achieve. But something of this notion has crept into the jargon of the trade with phrases like "error" rates, "false" negatives and "false" positives.

Alternatively, one might consider recidivism to be fundamentally a process involving chance with different individuals having different probabilities of failure. But it is recognized that the model is to provide information for a dichotomous decision. The pragmatic test of the model is the costs and benefits associated with the number of "right" and "wrong" decisions that would be made in practice.

---

[2]This over-simplifies the case. In the development of such a predictive instrument the researcher is inevitably interested in how its predictive accuracy compares with alternative decision schemes. The most obvious alternative is to assume that in terms of risk the subjects are indistinguishable. The recidivism probability assigned to all is the failure rate in the population as a whole. In this case, a measure of the benefit to decision making offered by a particular prediction device is the improvement over pure chance in the accuracy of a set of individual predictions, given the base rate for failures. (See, for example, Duncan et al. (1952), and Loeber and Dishion (1983)).

Finally, there exists a sub-class of studies, unfortunately rather small, in which a model is empirically validated by applying it to a population of subjects that is independent of the one used for model development. Most frequently, this is done under a split-sample design in which independent construction and validation samples are randomly drawn from the same population. This method is not without its critics (e.g. Maltz, 1984) and admittedly falls rather short of the scientific ideal of multiple, independent replications. Nevertheless, given the problem of saying something credible about generalizability based on a single sample design, an empirical test of predictive power, even within a particular study population, would seem to offer some essential information about how the model might perform in use.

In this paper we posit some simple assumptions -- not in the sense of theses to be tested formally but rather as points of departure determining what we attempt (and do not attempt) to do.

First, recidivism is assumed to be an intrinsically stochastic process with each individual having a characteristic failure probability.[3] The "stochastic" part of this assertion simply recognizes the impossibility of forecasting with any specificity the chain of events leading to recidivism -- no matter how recidivism is operationalized.

The "characteristic" part asserts that a person's habits and past experiences substantially determine the kinds of options leading to recivism that he is likely to encounter, the probabilities for the choices he will make and, perhaps, his chances of being observed to fail through his arrest or reconviction.

Second, individual failure probabilities are assumed to be changing in time. Typically, for the person who has not yet failed, one would expect these probabilites to be increasing for a time as he repeatedly encounters failure opportunities. But after awhile, resistance to such temptations itself becomes

---

[3]A corollary of this assumption is that each subject has a characteristic probability of <u>not</u> failing. Evidently, a model form that mathematically dictates eventual failure for all subjects is inconsistent with this assumption. An interesting line of research has, therefore, developed using mixed population models. Such models begin with the assumption that an observed population is made up of a sub-population of certain (or almost certain) successes and a complementary population of recidivists. In criminal career applications, this translates into sub-populations of desisters and persisters. (Maltz and McCleary, 1977; Harris et al., 1981; Maltz, 1984; Blumstein et al., 1985; Barnett et al., 1987.)

something of a habit and his risk of failure should begin to decrease.

Finally, we would assert that a model's performance under a split sample design contains information that can be used in an iterative procedure to improve its generalizability. For a number of reasons, a model can virtually never be expected to perform as well on new data as it did on the data used for model construction. The amount of degradation has come to be called "shrinkage" and statistical techniques are being developed for its estimation. (Copas, 1985; Copas and Tarling, 1986.). Given the assumption of a characteristic failure probability, what shrinkage effectively means is that any model is overly optimistic about its ability to discriminate between individuals based on the values of their independent variables. The empirical procedure explored in this paper is proposed as a method for reducing at least one of the major reasons for shrinkage.

The primary focus of this paper is methodological: to explore empirically the notion that selected actuarial models can "plausibly" uncover individual failure probabilities on the basis of bits and pieces of information about the offender and his history and that these models can be shown to have reasonable predictive validity -- at least under a split-sample test. Obviously, a failure probability is not a directly observable individual characteristic. Rather, it is a construct defined by the model itself. Given a model with its individually assigned failure probabilities, we make inferences about its accuracy by asking how plausible an observed pattern of successes and failures in a large population appears to be. Of interest is a comparison of the performance of mathematically different model forms -- in particular of the "static" logit model with a time dependent hazard model.

Finally, we investigate briefly some of the practical and theoretical conclusions to which we are led if we accept the models that have been developed.

A word about how the paper is organized. The next section deals in summary fashion with some of the mathematics of modeling used in the applications. Except for the definition of the particular form of the hazard model being explored here, this section will not be of much interest to people who either care little or know a great deal about such matters. Section III describes the general analytic plan followed in the investigations, whose results are reported in Sections IV and V. In Section IV the data being analyzed are taken from a two year follow up of male inmates released from prison in North Carolina. The data for the models of Section V cover ten years of criminal justice contacts following parole of the study subjects from a California Youth Authority institution.

Finally, as an Afterword, Section VI attempts to explain what the authors of this report think they have learned.

II. A BIT ON MODELING MATHEMATICS.

A. On Model Forms.

1. The Logit Model.

When the dependent variable is dichotomous (success=0; failure=1), the assumptions of linear regression models are violated and their use may give misleading results. (See, e.g., Aldrich and Nelson, 1984.) The non-linear model form most widely adopted to overcome this difficulty is the logit. This form is derived from the function

$$p(z) = \frac{1}{(1 + e^{-f(z)})}$$

If f(z) varies from minus to plus infinity, p varies from zero to one with symmetry through the point f(z)=0, p=1/2.

In the logit models investigated here, $f(\underline{z})$ is taken to be a simple linear function

$$f(\underline{z}) = \underline{z}'\underline{c}$$

where $\underline{z}'$ is to be understood as the transpose of a vector of values of k covariates measured on an individual (possibly including a constant) and $\underline{c}$ is a vector of k coefficients. Matrix multiplication is implied. Algebraically then,

$$\ln \frac{p}{1-p} = \underline{z}'\underline{c}$$

With p interpreted as the probability of failure within some fixed period of time, p/(1-p) expresses the odds of failing and this form of the logit model makes the assumption that the log of these odds for an individual characterized by $\underline{z}$ is simply a weighted sum of these variable values.

This is undoubtedly all familiar. Just as straightforward although less often discussed are the substantive interpretations to be given to the model parameters c_m. Standardized variables $\underline{z}$ do not seem appropriate for the kind of cross-validation investigations to be undertaken here so individual variables like age, sex, or number of prior arrests will cover very different ranges. This means that the values of the components of the vector $\underline{c}$ are not dimensionless and, consequently, are not directly comparable as measures of the contribution particular variables make to the recidivism probability.

But suppose there are two hypothetical individuals, i and j, who are identical in all respects except for their values on variable m -- their age, perhaps.

Then

$$\underline{z}(i)'\underline{c} - \underline{z}(j)'\underline{c} = (z_{im} - z_{jm})c_m$$

or in terms of failure odds

$$\frac{\text{Odds}(i)}{\text{Odds}(j)} = e^{(z_{im}-z_{jm})c_m}$$

These expressions involve only the one parameter c_m and its associated variable values $z_{.m}$. It is in this sense that the "strength" of the individual variables will be isolated and reported here.

While this is mathematically correct, a word of caution about the phenomenological interpretation of these comparisons might not be out of order. A model is a sort of "recidivism gestalt," built out of a particular set of facts and attributes selected in advance by the analyst. The role of any one variable cannot be divorced from the analytic context of all independent variables appearing in the model. Consider, for example, a variable such as "length of present term of incarceration." This variable conceivably could show very different "strengths" in models that do or do not also include the variable "age at release" if, in fact, the subjects in a study population who served longer terms also tended generally to be the older ones. The interpretation of "strength" adopted here is the usual one that appears in the literature as "controlling for other factors." The caution is simply that this can have no real meaning without an understanding of what those other factors are. In particular, it can at this stage of criminology's development never be read as "controlling for all other factors."

2. Hazard Models.

Criminology was introduced to hazard models by Stollmack and Harris (1974). They were interested in the idea that time to failure might be a more sensitive measure to apply in a program evaluations than the more traditional failure rate within a specified follow-up period. Notable among the subsequent developments and applications of this methodology in recidivism studies are the works of Schmidt and Witte (1980), Barton and Turnbull (1981), Schmidt and Witte (1984), Maltz (1984), and Schmidt and Witte (1987).[4]

---

[4]For a more thorough mathematical exposition, see Lee (1980) and Kalbfleisch and Prentice (1980) -- especially for a discussion of biased censoring mechanisms and application of the proportional hazard model with a non-parametric time dependence.

The hazard function h(t) is defined as the conditional probability per unit time that an individual who has survived to time t will fail in the time between t and t+dt. Equivalently, in a very large population it is the rate at which failures are occurring at time t among those subjects who have survived to t.

From this definition it follows that, if S(t) is the unconditioned probability of survival to t, then S(t)h(t)dt is the unconditioned probability of failure in the interval t to t+dt. In frequentist terms S(t) is the expected fraction of the initial population surviving to time t, S(t)h(t)dt the fraction of the initial population failing in [t, t+dt]. Therefore,

$$dS = -S(t)h(t)dt$$

and it follows that

$$S(t) = e^{-\int_0^t h(x)dx}$$

Researchers, such as those whose works are cited above, have investigated the implications of a variety of mathematical forms for the hazard function h. Here we are interested in finding a form consistent with the assumption that the typical subject's hazard rate begins by increasing in time, passing through a maximum and then decreasing.[5] The form considered in this paper is

$$\ln h(t,\underline{z}) = \underline{z}'(\underline{c} + \underline{b}t + \underline{a} \ln[t])$$

or

$$h(t, \underline{z}) = e^{\underline{z}'(\underline{c}+\underline{b}t)}\, t^{\underline{z}'\underline{a}}$$

Here, again, $\underline{z}'$ is the transpose of a vector of measurements of k

---

[5]There are, of course, many functional forms that would satisfy these conditions as well as the obvious requirement of being non-negative for all positive t. The form adopted here is, perhaps, the simplest at least in its property of being loglinear in the covariates. It might also be noted that the approach here differs from that of some researchers who have investigated parametrized forms of hazard models in that it makes an a priori assumption of the form of the hazard function rather than of the probability density. As pointed out in the Note following this section, this can lead to a defective probability distribution, and thus introduce some statistical complications. For example, the expected time to failure for any class of subjects is then defined only conditionally: given that a failure will occur.

variables made on each subject and $\underline{c}$, $\underline{b}$, and $\underline{a}$ are each vectors of k parameter values -- some of which may be constrained to be zero.

Some mathematical consequences of the assumption of this form for the hazard function are discussed in the Note at the end of this section. Here the concern is the substantive interpretation of the strength of the contribution each of the variables makes to a subject's recidivism probability.

Just as in the logit model discussion, assume there are two individuals i and j differing only in their values on the m^{th} variable. The log of the ratio of their hazard functions at time t is then

$$\ln\left(\frac{h_i(t)}{h_j(t)}\right) = (z_{im} - z_{jm})(c_m + b_m t + a_m \ln[t]).$$

This ratio, of course, compares the probability of failure in the near future of subject i to that of subject j -- assuming that both have survived to time t. Again it is in this sense that the contributions of the different variables will be compared. Unlike the comparison of odds in the logit model, however, the relative strengths of different variables will be changing over time. Some mention might be made of how the values of a_m, b_m and c_m determine this change.

Consider the function

$$g_m = e^{(c_m + b_m t + a_m \ln[t])}.$$

For very small t, ln(t) will be negative and very large in absolute value. This term will dominate ln g_m. If a_m is positive, g_m will approach 0 as t approaches 0. Conversely, g_m will tend to infinity near t = 0 if a_m is negative. For very large t, the term $b_m t$ is dominant so that g_m eventually tends to infinity or zero, depending on whether b_m is positive or negative.

It might be noted in making the comparisons between subjects i and j that the quantity $(z_{im} - z_{jm})$ can be positive or negative, depending on which subject is arbitrarily designated as i. The limits of the hazard ratios at t = 0 and t = infinity will be zero or infinity, depending on this choice. All this means, of course, is that the interchange of subject i and j inverts the ratio of hazards so that zero maps into infinity and vice versa. It should be emphasized, however, that, if the variable z_m can take on only positive or only negative values for any subject, the "partial hazard"

$$h_m = e^{z_m(c_m + b_m t + a_m \ln(t))}$$

has the same limiting values at zero and infinity for all subjects.

Roughly speaking the parameter a_m gives information about how the variable z_m influences the recidivism probability over the very short term following release; b reflects its long term influence given long term survival. But a bit of care must be exercised in interpreting the h_m ratios at very small and very large t. A very large value of this ratio near t = 0, for example, does not mean that subject i is doomed to instant failure. For both i and j the conditional probability of failure may be very small in an absolute sense; but in a relative sense i's probability is much greater than j's in the early days of time at risk.

If z_m is a categorical variable taking only values zero or one, the interpretation is similar: the ratio is a comparison of the partial hazard of the subject with $z_m = 1$ to the fixed value $h_m = 1$ that is assigned <u>a priori</u> to the subject with $z_m = 0$.

The parameter c_m plays a rather complicated role in that it determines a value about which h_m varies in the course of time. Its function in the hazard model is discussed in somewhat greater detail in the Note following this section. Here it will suffice to mention two rather obvious roles it can play.

If both a_m and b_m are zero, the partial hazard h_m is evidently constant in time with the value $z_m c_m$. Thus, the variable z_m enters as a constant multiplicative factor $\exp(z_m c_m)$ into the log of the survival probability at time t. But in the more general case, consider the equation

$$c_m + b_m t + a_m \ln(t) = 0.$$

Depending on the values of c_m, b_m and a_m, this equation can have 0, 1 or 2 roots in t. In terms of the comparisons between subjects i and j, the roots of this equation are the times at which these two subjects are at equal risk. Their relative characterization as being the one at higher or lower risk then switches as t passes through a root value.

B. On Likelihood Estimation.

Given a model form with a particular set of parameter values, each subject in a study is assigned an individual probability of failure -- failure before some fixed time T in models like logit or failure within a variable interval t to t+dt in hazard models. By definition the likelihood function is simply the joint probability of occurrence of the whole pattern of observed

failures and successes, given the individual, model-assigned failure probabilities. The log of the likelihood function is, thus,

$$\ln L = \sum_{\text{failures}} \ln(p_i) + \sum_{\text{successes}} \ln(1-p_i)$$

for a fixed-time model. For a hazard model the pattern of observations to be modeled includes not only the outcome variable, success or failure, but also the time at which failure occurred for subjects who failed and the time of censoring for subjects who did not. The unconditioned failure probability for subject i is, again, $S_i(t)h_i(t)dt$. Dropping the constant factors dt, the log likelihood is

$$\ln L = \sum_{\text{failures}} (\ln h_i[t_i] + \ln S_i[t_i]) + \sum_{\text{successes}} \ln S_i[t_i]$$

or

$$\ln L = \sum_{\text{failures}} \ln h_i(t_i) + \sum_{\text{all}} \ln S_i(t_i).$$

Here t_i is the observed time to failure or censoring of subject i. Since

$$S_i(t_i) = e^{-\int_0^{t_i} h_i(x)\,dx},$$

it follows that

$$\ln L = \sum_{\text{failures}} \ln h_i(t_i) - \sum_{\text{all}} \int_0^{t_i} h_i(x)\,dx.$$

Given competing sets of parameter values, it is certainly reasonable to prefer the one that would assign the greatest joint probability to the outcome pattern of failures and successes actually observed. The set of parameter values with the greatest probability among all possible sets is, of course, the maximum likelihood estimate of the "true" values. It is an estimate in the usual statistical sense in that it derives from a particular data base that may not be in every detail a faithful copy of the underlying population it is assumed to represent.

The mathematical problem of likelihood maximization reduces to finding a solution to the system of equations obtained by setting to zero the first derivatives of the log likelihood with respect to each of the model parameters -- thus, with k parameters, a system of k simultaneous, non-linear equations.

In the computational process of carrying out a systematic, iterative search for solution values, the set of k(k+1)/2 second

derivatives of the log likelihood are also evaluated at each step. If

$$J_{ml} = \frac{\partial^2 \ln L}{\partial c_m \partial c_l} ,$$

then

$$\underline{C} = (-\underline{J})^{-1}$$

gives an estimate of the covariance matrix of the parameters. Thus the square roots of the diagonal elements of $\underline{C}$ are estimators of the parameters' standard deviations.

Note on the Gamma Hazard Model.

The hazard model investigated in this paper has the form

$$\ln h(\underline{z}_i,t) = \underline{z}_i'(\underline{c} + \underline{b}t + \underline{a}\ln(t)).$$

Here $\underline{z}_i$ is a vector of measurements of k variables made on subject i; and $\underline{c}$, $\underline{b}$, and $\underline{a}$ are each vectors of k parameters -- some of which may be constrained to be zero. Matrix multiplication is implied: the right hand side is a scalar function of t.

This particular form was chosen because, if $\underline{z}'\underline{a}$ is greater than zero and $\underline{z}'\underline{b}$ less than zero, the resulting hazard function starts out at zero at t = 0, increases to a maximum at

$$t_{max} = \frac{\underline{z}'\underline{a}}{|\underline{z}'\underline{b}|}$$

and then again decreases to zero as t becomes indefinitely large. This seems a mathematically simple way to reflect a plausible *a priori* assumption about how the probability of recidivism might change over time, conditioned on no recidivism up to time t.

With this form of the hazard the value at maximum is

$$h_{max} = \left(\frac{\underline{z}'\underline{a}}{|\underline{z}'\underline{b}|}\right)^{\underline{z}'\underline{a}} e^{\underline{z}'(\underline{c}-\underline{a})}.$$

Furthermore, given any specified positive value h_o, the equation

$$\underline{z}'\underline{c} + (\underline{z}'\underline{b})t_o + \underline{a}\ln(t_o) - \ln(h_o) = 0$$

has either 0 or 2 roots t_o. If there are no roots, the hazard function never reaches the value h_o. In the case of 2 roots, the value of the hazard is greater than h_o during the time interval between them.

Thus, with this form of hazard function, the parameter vectors $\underline{c}$, $\underline{b}$, and $\underline{a}$ would allow the individual subject's covariates to determine the time at which he is at greatest risk of failure, the magnitude of the risk, and its spread in time -- the last in the sense of the time interval over which his risk is greater than any specified value.

Of course, the hazard function can assume other shapes than that of a typical "gamma density". In particular,

1. It is monotonic in t if $\underline{z}'\underline{a}$ and $\underline{z}'\underline{b}$ have the same sign (increasing if they are both positive, decreasing if negative); and

2. It is U-shaped if $\underline{z}'\underline{a}$ is negative and $\underline{z}'\underline{b}$ positive.

The probability of survival to time t is, as usual, given by

$$S(\underline{z},t) = e^{-\int_0^t h(\underline{z},x)\,dx}.$$

The integral does not exist unless $\underline{z}'\underline{a}$ is greater than -1. This could create something of a nuisance in the analysis. In fact no problems have been encountered in parameter estimation. Under cross-validation, models estimated on different data bases have sometimes assumed invalid forms. When this situation arises, it is most often the case that the "a" coefficients of a particular variable have taken on relatively widely separated values in the two models. For some of the study subjects (no more than one or two in our experience) this has resulted in an invalid form of the model on cross validation.

One property of interest is that, if the integral exists and if $\underline{z}'\underline{b}$ is negative, the integral converges in t. This means that under these conditions the model dictates a finite probability of long-term survival:

$$\ln S(\underline{z},\infty) = -\int_0^\infty h(\underline{z},x)dx = e^{\underline{z}'\underline{c}}\frac{\Gamma(\underline{z}'\underline{a}+1)}{|\underline{z}'\underline{b}|^{(\underline{z}'\underline{a}+1)}}$$

where

$$\Gamma(\underline{z}'\underline{a}+1)$$

is the gamma function.

Obviously, it would be very unsound to use such a model to project survival probabilities far beyond the span of time of the observations on which the model is built. At the same time this analytic feature of the model (or of any hazard model for which the integral of h(x) converges as the upper limit becomes indefinitely large) might be considered of some interest in that its form does not imply a prior assumption that all subjects must eventually recidivate.

A logically consistent interpretation of such a "defective probability distribution" is to consider the recidivism model as part of a larger model in which, over the long run, everyone must indeed fail in one way or another. That larger model would then consider as competing risks all the possible events a subject might experience whose prior occurrence would preclude the possibility of failure through recidivism. A subject's death is perhaps the most obvious example. In this sense the probability of survival to time t being investigated here is clearly a

conditional probability -- the probability of no recidivism before time t, given that the subject is actually at risk of recidivating over the whole interval [0, t].

The assumption of a log linear form in the covariates means that the hazard function is a product of "partial hazards", each of which is a function of time and a single covariate:

$$h_m(z,t) = \exp[z_m(c_m + b_m t + a_m \ln(t))],$$

where z_m is a subject's measure on the m^{th} covariate and c_m, b_m and a_m are the m^{th} components of the coefficient vectors $\underline{c}$, $\underline{b}$, and $\underline{a}$ respectively. This property simplifies considerably the interpretation of the relative strengths of the different covariates in their contribution to recidivism probabilities.

Finally, it should be noted that the values of the coefficients will depend on the unit of time chosen in the estimation. The transformation from one time scale to another is straightforward. Suppose, for example, time t is measured in units of years but it is desired to express everything in units of months. That is, coefficients of the model in which t"=12t are wanted. More generally, consider the time scale transformation t" = rt, where r is a constant.

Since the quantity h(t)dt is a conditional probability, it must be invariant under this change so that

$$h"(t")dt" = h(t)dt.$$

Substituting t=t"/r, one readily obtains

$$\exp[\underline{z}'(\underline{c} + \underline{b}t + \underline{a} \ln[t])]dt$$

$$= \exp(-\ln[r] + \underline{z}'([\underline{c} - \underline{a} \ln(r)] + (\underline{b}/r)t" + \underline{a} \ln[t"]))dt".$$

Let the first component of each of the coefficient vectors be the constant term. Then the coefficient values for the model with time measured in the units of t" are related to the originally estimated coefficient values by

$$c_1" = -\ln(r) + c_1 - a_1 \ln(r);$$

$$c_m" = c_m - a_m \ln(r) \qquad m \text{ not equal to } 1;$$

$$b_m" = b_m/r \qquad \text{all } m;$$

$$a_m" = a_m \qquad \text{all } m.$$

III. THE ANALYTIC APPROACH.

A. The Overall Scheme.

The two basic problems of building an actuarial model of recidivism are determining which covariates are systematically related to observed failure and deciding how these variables are to be combined mathematically to produce an estimate of failure probability.[6] Given a solution to the first problem, the second would reduce to investigating the relative validity of different model forms -- logit versus hazard models, for example. But in the absence of strong, well-confirmed theory of what individual attributes lead to success or failure, the analyst evidently faces something of a dilemma.

The persistence of this dilemma over years of criminological research can, at least in part, be traced to the fact that quite independent theoretical concepts cannot easily be translated into statistically independent observables. Suppose measures of k covariates have been made on a subject population. Presumably, each of these was chosen because it has a theoretically plausible relation to recidivism and, consequently, cannot be rejected <u>a priori</u>. It is virtually certain that there exists a fair amount of correlation between these variables. For example, age at release, length of criminal record and time served in the last incarceration all have independent, theoretical justification as predictors of recidivsm. But length of criminal record is likely to increase with age across the population (for any given subject, it certainly cannot decrease); age at release might be expected to increase with time served; and time served probably increases with length of record. Which of these three or which combination actually has the greatest validity as a predictor must ultimately be determined empirically.[7]

---

[6]Discussion of the controversy of how "failure" itself should be operationlized is beyond the scope of this paper. We follow Maltz (1984) in using arrest for a new crime as the least objectionable among the alternatives -- at least for the purposes for which these analyses are undertaken.

[7]Two systematic approaches that address the problem of which of a set of available covariates to include in a model are to be found in the recidivism literature. The first is simply to re-estimate the model after removal of covariates found not to be significant at some pre-specified level. The second is to use multi-stage regression or some equivalent procedure to select the covariate that would at each step make the greatest improvement in model's fit. In terms of improving predictive power, however, both methods are open to criticism. (Copas (1985))

Implicit in the assertion that one model is demonstrably better than another is the notion that there exists an empirical scale for measuring how well any given model performs when applied to a given population sample. A number of analytic measures might be proposed that derive from the likelihood function. Since the likelihood is, by definition, simply the joint probability of occurrence of the observed pattern of failures and successes when individual failure probabilities are prescribed by a given model, it provides a straightforward measure by which to compare the relative explanatory power of different models. But it has some limitations when used as a basis for a statistic that attempts to provide an absolute measure in some "goodness of fit" sense.

A simple, patently contrived example might serve to illustrate the point.

Suppose in a study population of 1000 subjects, 550 failures and 450 successes are observed. Under a naive model, all subjects are assumed to be the same and the individual failure probability is taken to be the population failure rate, .55. The likelihood under this model is

$$L_o = (.55)^{550} \times (1-.55)^{450}$$

or

$$\ln L_o = -688.1.$$

Suppose under a model that relates failure to a single, categorical covariate (and includes a constant term), 500 subjects are assigned a failure probability of .50 and 500 a probability of .60. Suppose further that there were 250 observed failures in the first group and 300 in the second. The log likelihood in this case is given by

$$\ln L = 250 \ln(.5) + 250 \ln(.5) + 300 \ln(.6) + 200 \ln(.4)$$

$$= -683.1.$$

There is thus some modest improvement in explanatory power under the second model. Indeed, the likelihood ratio test in this case gives a chi-square of 10, [2(688.1-683.1)], which is statistically significant at a .005 level with 1 degree of freedom.

While this suggests that the covariate of the hypothetical model does indeed bear a relationship to recidivistic failure, one might well want some absolute measure of how powerful this relation is. A rather straightforward measure of this might be the Nth root of the likelihood, where N is the total population size. This, of course, is just the geometric mean of the individual outcome probabilities: the model-assigned failure probabilities of those who failed and the success probabilities of those who did not. In this example the value is .505, which

is a slight improvement over the naive model's .503.

The problem with using statistics based on the likelihood as absolute measures of the power of a model is that in a sense it forces one towards a deterministic conception of the recidivism process. If failure or success were the inexorable outcomes for individuals with or without certain traits, the likelihood might logically be seen as a scale for measuring how well a given model with a given set of covariate measures captures this elusive causal mechanism. The perfect model would predict individual futures with probabilities of 1.

This kind of determinism may be an appropriate basis for interpreting results of studies in biology or medicine but one would be hard put to find empirical support for it in studies on criminal recidivism. Covariates such as age at first arrest or number of prior convictions may well be seen as measures of a variable propensity to recidivate. But suppose (as we do here) that chance plays an inherent role in the outcome and consequently in our ability to predict the outcome. It's as if success or failure were being determined by an "unfair" coin toss with each study subject tossing his own peculiar coin. The job of the model then is not to decide whether subject j does or does not have the mark of Cain but to try to figure out the probability that j's coin will come up heads.

If that's the case, a measure of overall "goodness of fit" such as the N^{th} root of the likelihood will almost certainly remain disappointingly in the .50 -.60 range, no matter what marginal improvements we make in the model. Such a measure is simply telling us that the model really doesn't do very well in making absolute assertions about who will fail and who will succeed.

To get back to the example, a probabilistic view of the recidivism process might interpret the accuracy of the model by noting that, with a failure probability of .60, the expected number of failures in a group of 500 identical subjects is 300 with a standard deviation of 11.0. For the group of 500 with failure probability .50, 250 failures would be expected with a standard deviation of 11.2. Since the number of failures observed in the two groups of this contrived example were 300 and 250, respectively, the actuarial-minded might indeed take some satisfaction in the model's ability to fit the data. But it should also be noted that the naive model does just as good a job in this sense since 550 failures would be expected among 1000 subjects if they uniformly had an assigned failure probability of .55.

The reason for this rather tedious discussion of what must seem quite obvious is that it is fundamental to the approach adopted in this inquiry. The likelihood function is the basis for deciding if one set of covariates is better than another in

capturing statistical relationships of importance to the individual failure probabilities. Statistics based on the binomial distribution are the measures used to rate the goodness of the fit of any given actuarial model.

B. Model Construction.

Parsimony is generally advocated as one of the basic principles of good procedure in model building. (e.g. Box and Jenkins, 1976:17). A model built on too rich a set of explanatory variables will be correspondingly weak as a predictor when applied to new data. Copas (1985) illustrates this point by suggesting using subjects' names as variables. A recidivism model judiciously built on such "data" could be made to fit perfectly. But it would scarcely be expected to have any predictive power.

We assume that there is some structural relation between an individual's probability of recidivism and his measured values on some vector of theoretically defensible covariates. Further, we assume that such a structure is discoverable from observations made on a subject population. The problem is that we must expect random inter-sample differences in the correlations among variables and in their relationship to failure. Thus, to some unknown extent, each population sample is idiosyncratic. Shrinkage in the power of a set of independent variables to explain recidivism is to be expected when the fitted model is validated on a different population sample. The more variables used to achieve precision in explaining relationships found in the construction data, the greater the tendency of the model to reproduce faithfully that data's quirks and eccentricities. (See e.g. Reiss, 1951). Obviously such a model would not fit very well on data that has a randomly different set of peculiarities. The question is how to tell the difference between structure and noise.

Given a set of candidate independent variables, this part of the modeling problem can be defined as determining which subset of variables can in some sense be regarded as doing the "best" job of simultaneously capturing the maximum of structural information while at the same time modeling a minimum of sample noise.[8] A

---

[8]Note that this search is antecedent to the problem of estimation of and correction for the shrinkage expected from a given model built on a given data set. Criminological applications of statistical methods for shrinkage estimation are given in Copas (1985) and Copas and Tarling (1986). For an information theoretic approach see Larimore (1983) and Larimore and Mehra (1985).

one-sample procedure analogous to multi-stage regression could be used with models estimated by likelihood maximization. The likelihood ratio test would provide a measure of the statistical significance of the increase in likelihood produced by each variable added to the model. With considerably less statistical elegance and a correspondingly greater reliance on empirical results, we proceed here somewhat differently.

The analytic plan follows this general recipe:

1. The study population is randomly divided into three non-overlapping samples of approximately equal size. Call them samples A, B and C.

2. Starting with a model that includes all independent variables of potential interest, coefficients are separately estimated on data samples A and B.

3. The model estimated on B is applied to sample A data and vice versa. The four log likelihoods (two from estimation of the models on A and B and two from cross-validation) are then added. (Hazard models are treated here just like fixed time models. That is, the model parameters are used to calculate for each subject the probability of failure before some common time T -- two years, perhaps.)

4. Constraining each of the coefficients in turn to be zero, approximations to the change that would be found in this sum of log likelihoods are now calculated. This, of course, is an attempt to approximate what the results might look like if one independent variable were deleted from the model.

5. The variable whose elimination would result in the greatest algebraic increase in the sum of log likelihoods is dropped and the analysis starts over with the reduced set of independent variables.

6. This programmed search ends when, in this approximation, dropping any of the remaining variables would decrease the log likelihood sum.

7. If the final model is thought to be acceptable, the coefficients are then estimated on sets A and B combined and the "predictive" accuracy of the model is tested by validation of the result on data set C.

Before setting out the mathematics of this estimation process, it might not be out of place to give a plain language description and, hopefully, some justification for what is going on at each step.

In steps 1 and 2 it is asserted that each of the three randomly

selected sub-samples of the data has an equal claim to being a faithful copy of the underlying population of which the whole data base is itself a particular sample. (Somewhat more precisely, the claim is an increasing function of the number of subjects in each sub-sample. We take for granted that these sub-samples are large enough in relation to the number of model parameters to give reasonable assurance that estimates have a chance of being stable.) We, therefore, have no reason to prefer either the A or the B model. This is the underlying reason for summing the four log likelihoods in step 3.

Consider for the moment the model built on data set A. The two log likelihoods produced by this model (on A and B data separately) are independent probabilities whose product is the joint probability of observing the pattern of outcomes found in the combined sample -- conditioned, of course, on accepting model A's view of the world. The two likelihoods produced by model B are interpreted similarly. This means, of course, that we now have two competing views of the same world. To what extent can they be reconciled?

By adding the four log likelihoods, we obtain a function which, except for omission of a factor of 1/2, is the log of the geometric mean of these world views. We take this to be a simultaneous measure of the discrepancy between them and the firmness with which each view is held. The former is measured by the two cross-validation terms, the latter by the construction log likelihoods.

Since we are concerned at this stage of model construction with the parsimony issue, step 4 tries to arbitrate between the two models by proposing to eliminate one of the independent variables. Any such elimination must, of course, decrease both of the construction likelihoods. The amount of the decrease in each case measures the importance the model attaches to that variable in its explanation of its construction data results.

It is possible, however, that dropping a variable could increase the cross-validation likelihoods. (For example, we would expect this to be the case if the coefficients of a particular variable have different signs in the two models.) We take such an increase as an indication that inclusion of the variable in question might be an important factor contributing to the discrepancies between the models' world views.

At step 5 a compromise is imposed. That variable is selected whose elimination would on balance produce the greatest benefit in terms of reduction in the discrepancies between the two models net the cost of poorer fits to the two sets of construction data.

The process is repeated until no further compromise can be reached through elimination of more variables. Step 7 then

dictates a best estimate of a single world view based on the combined data sets and those explanatory variables not eliminated in this process.

What follows is the mathematical description of the process as it is applied in the analyses of this paper.

Let L(Y,x) denote the likelihood calculated on data sample Y with the paramaters of a model estimated on sample X. Define

$$U(A,B) = \ln L(A,a) + \ln L(B,a) + \ln L(B,b) + \ln L(A,b).$$

If the discussion is restricted for the moment to fixed time models such as the logit, the terms L(A,a) and L(B,b) are the values of the likelihoods obtained in the model estimations, which means that for a given set of variables and a given data sample the parameter sets a and b are picked so that they maximize these likelihoods. Any change in parameter values, in particular the imposition of a constraint that the coefficient of one of the variables be zero, must decrease these likelihood values and, of course, their logs. Suppose, however, that a particular variable bears a quite different relation to recidivism probabilities in samples A and B. Elimination of this variable from the model might then produce a substantial increase in the cross-validation log likelihoods. It is the net change in U when any given parameter is set to zero that is of interest here.

The argument is virtually identical when hazard models are being explored. In this analysis the choice was made to calculate the function U by computing the individual success and failure probabilities and the four associated log likelihoods at a fixed time, thus allowing comparison of the results derived from mathematically different model forms. Model estimation using likelihood maximization with hazard models is somewhat different than with fixed time models in that the likelihood function depends not only on which subjects failed and which succeeded but also on the times at which the failures occurred or the times at which observation of the successes was stopped. It cannot, therefore, be asserted that the terms ln L(A,a) and ln L(B,b) would necessarily decrease under any change in parameter values since these are not precisely the functions that were maximized in the estimation procedure. In fact this makes no difference to the argument in support of the change in U as a plausible guide for exploring the parsimony question.

The changes in U could obviously be calculated directly by re-estimation of a model reduced by one variable. This would be a very time consuming procedure since at each stage of the exploration the process requires separate testing of each of the variables still under consideration. To speed things up, the

changes are approximated as perturbations, using the first few terms of a Taylor series expansion of U. Thus, with k coefficients we write for the change in U when the r^{th} coefficient is set to zero

$$\delta U_r = \sum_{j=1}^{k} \left(\frac{\partial U}{\partial a_j} \delta a_{jr} + \frac{\partial U}{\partial b_j} \delta b_{jr} \right) + \frac{1}{2} \sum_{j=1}^{k} \sum_{m=1}^{k} \left(\frac{\partial^2 U}{\partial a_j \partial a_m} \delta a_{jr} \delta a_{mr} + \frac{\partial^2 U}{\partial b_j \partial b_m} \delta b_{jr} \delta b_{mr} \right)$$

The "r" subscript is simply a reminder of which coefficient is being constrained. The notation is not explicit but it is to be understood that all derivatives are calculated with current values of the coefficients. From the definition of U all second order mixed partials with respect to an "a" and a "b" coefficient vanish.

We next have to determine the approximate changes $d\underline{a}$ and $d\underline{b}$ in the remaining coefficient values when the r^{th} coefficient is set to zero. The values of the coefficients were determined by solving with A and B data separately the k simultaneous, non-linear equations resulting from setting the first partial derivatives of the log likelihood equal to zero. Consider for the moment just the "a" coefficients. Let $V_i(\underline{a})$ denote the first derivative of the A data log likelihood with respect to a_i. $V_i(\underline{a})$ is to be considered as a function of all k "a" coefficients. (The likelihood function here is the form used in model estimation. For logit models it is the same form as the functions entering into the definition of U: ln L(A,a). But for hazard models the functional forms are different in U and V_i because of the way we have chosen to define U). We now expand these k functions in a Taylor series about the solution values:

$$V_{ir}(\underline{a}+\delta\underline{a}_r) = V_i(\underline{a}) + \sum_{j=1}^{k} \frac{\partial V_i(\underline{a})}{\partial a_j} \delta a_{jr}$$

Again, the r subscript is simply a reminder that the formalism aims at determining the effect of elimination of a particular variable. And again, it is to be understood that the functions on the right are evaluated with the current values of the parameters. Consequently, for any r the first term on the right vanishes in each of the expressions i = 1, ... k. To approximate what happens when the r^{th} coefficient is set equal to zero, we now specify that

$$\delta a_{rr} = -a_r$$

We approximate the maximum likelihood solution for the parameter values of the reduced model by setting V_{ir} equal to zero for all i except i=r:

$$\sum_{j=1}^{k} \frac{\partial V_i}{\partial a_j} \delta a_{jr} = 0. \qquad i = 1, 2, \ldots (r-1), (r+1) \ldots$$

For each r value this produces a system of (r-1) linear, inhomogeneous equations. Along with the condition $da_{rr} = -a_r$, this determines the set of da_{jr} to be used in the calculation of dU_r. An identical procedure, of course, is used to determine the set of db_{jr}.

Parenthetically, it might be noted that

$$\frac{\partial V_i}{\partial a_j} = \frac{\partial^2 \ln L}{\partial a_i \partial a_j}$$

The matrices of coefficients in the equations determining the da_{jr}'s and db_{jr}'s are thus related to the estimators of the information matrices and to the estimated covariance matrices of the parameters. Mathematically, the procedure of model investigation outlined here is, therefore, not unrelated to the traditional arguments of variable significance based on estimations of parameter t-statistics. It differs, of course, in its adoption of a decision rule for "non-significance" that is in part based on empirical results deriving from the simultaneous estimation and cross-validation of models using two independent sub-samples of observations.

The values of dU_r are calculated for each r from 1 through k. Positive dU values indicate, at least in this approximation, that U would increase if any of these variables were dropped. As argued above, this is taken as evidence that a more parsimonious model would eliminate some of the differences in the world views that exist between the separately estimated A and B models. Given any positive dU values, the decision rule, of course, is to drop the variable associated with the largest. The whole process, beginning with separate parameter estimations on A and B, is then carried out with the reduced set of independent variables. This iteration procedure ends when, for a given set of variables, all the dU_r turn out to be negative. Parameter estimates for a "final" model are then obtained from the combined A and B data.[9]

This procedure can lay no claim to providing a unique solution to the problem of finding that model that "best" fits both of the data samples. It is devised as a systematic and plausible substitute for the impossible task of exhaustive exploration of all of the 2^k-1 non-empty sub-sets of variables that might be

---

[9]Note that it is at this point that one could introduce statistical methods for obtaining "pre-shrunk" parameter estimates as in Copas (1985).

selected as candidate models.

In the course of the analysis the algorithm may make some decisions that are questionable for one reason or another and so should be explored further. Perhaps the simplest reason for the procedure to take a turn down a dubious path is the truncations of the Taylor series that are explicitly built into the procedure. For example, the da_{jr}'s that result from setting da_{rr} equal to $-a_r$ may not be "small" enough so that the first few terms of the series give a reasonably good approximation to the values of the functions at the point $\underline{a} + d\underline{a}_r$. This can be monitored to some extent by noting whether the value of U at the next iteration is reasonably close to the value expected from the approximation.

More problematic is the situation in which the procedure throws out what looks like a perfectly good variable -- that is, one in which the coefficients in the two estimated models seem significant by their t-statistics and approximately equal in value. This can happen even when the truncated Taylor series proves to be a good approximation to dU. So by the decision rules of the algorithm it is a legitimate move.

The reason for this rather odd behavior is that parameter values depend on the whole constellation of variables currently appearing in the model. If two of the explanatory variables, x and y, are correlated to some degree (or more generally, if x is substantially correlated with some sub-set of independent variables), it is possible that the removal of x may change the remaining parameter values in such a way that the improvement in cross-validation log likelihoods overcomes the self-validation log likelihood decreases. But again, the question of whether the algorithm made a right decision can best be answered "empirically" -- perhaps, by seeing what happens when variable x is re-introduced into the model at the end of the programmed search procedure.

C. On Model Validation.

A number of analytic measures might be proposed that derive from the log likelihood obtained when a model is validated on a data sample other than the one used for estimating its coefficients. But a test that is perhaps somewhat more closely akin to the actuarial spirit that underlies this investigation would seem more appropriate to a judgment of how accurate the model might prove as a predictor.

For each individual in a sample, the outcome, failure or success, is like the result of tossing an "unfair" coin. There are only two possibilities with p(i) being subject i's chance of failure, 1-p(i) his chance of success. In a large population of subjects,

all of whom are identical to subject i, the probability of finding any given number of successes and failures would follow a binomial distribution. On the average in a population of size N, one would expect to find Np(i) failures with a variance of Np(i)(1-p(i)).

Generalizing this, suppose we consider all those subjects in a sample whose model-assigned probabilities lie within some specified range, say p_1 to p_2. Out of this group the expected number of failures is just the sum of the individual probabilities

$$E(nf) = \sum_{p_i \epsilon (p_1, p_2)} p(i)$$

and the variance the sum of p(i)(1-p(i)):

$$\sigma = \sqrt{\sum_{p_i \epsilon (p_1, p_2)} p(i)(1-p(i))}$$

Suppose then that in examining the cross-validation results the overall range of model-assigned probabilities is divided into a number of segments. Let $E(nf)_k$ and Ob_k be the expected and observed numbers of failures in the k^{th} segment and n_k the number of subjects that the model assigns to that segment. Then the quantity

$$\sum_k \frac{n_k (E(nf)_k - Ob_k)^2}{E(nf_k)\ (n_k - E(nf_k))}$$

may be assumed to be chi-square distributed with degrees of freedom equal to 2K-1 where K is the total number of segments into which the range was divided.

This follows from the fact that both the n_k and the $E(nf)_k$ are determined by the model itself with the single constraint being that the sum of the n_k must equal the number of subjects in the population. This is somewhat different from the more familiar examples of application of the chi-square test, in which the classification of subjects into categories is determined on the basis of some observable characteristic like sex or age group.

Here the null hypothesis is evidently that the observed and expected numbers of failures over the set of K intervals are from the same distributions. This chi-square test thus gives a familiar statistical measure of the plausibility of the model as demonstrated under a cross-validation. The analyst may derive further information, of course, from an examination of the standard deviations. Certainly, it would be quite encouraging to find that, for all k, the absolute value of $E(nf)_k - Ob_k$ is less than the traditional 1.96 times the estimated standard deviation for that interval.

While this chi-square test gives a simple and readily

intelligible result, it does have certain drawbacks. The first is not really serious although it adds a bit to the computational problem. For finite sized populations, the probability associated with a chi-square value depends on the validity of an approximation that requires that all cells be adequately populated. The usual rule of thumb is that no cell contain fewer than 4 or 5 subjects. In the application here, this means each of the K divisions of the probability range must have at least this number of expected failures and expected successes. At the low end of the model-determined probability range this can mean combining the first few categories into one in order that the sum of assigned p-values meet the minimum requirement. At the high end of the p-range the problem is the analogous one -- in this case, getting enough expected successes. But, again, this is more of a nuisance than a troublesome flaw in the interpretation of the result as a measure of plausibility.

More serious, perhaps, is the problem that the test depends on the arbitrarily chosen number of categories K and therefore offers something less than an ideal measure of goodness of fit. If K is chosen too small, the measure is unsatisfactory in that it doesn't tell us much about how precisely the model has distributed the failure probabilities among the subjects. But by the same token, if K were taken to be very large and at least one of the independent variables is continuous, each of the segments of the probability range would ultimately have either 0 or 1 subject -- which is equally uninformative besides being about as gross an error as one could make in using a chi-squared statistic. The reason for this complication is that categorical variables necessarily introduce a certain lumpiness into the assignment of probabilities over the population. With M variables each able to take only 1 of two possible values, the subjects are necessarily distributed among the 2^M distinct vectors that could be constructed from these values. If a model eventually becomes quite parsimonious, with a few continuous and a few categorical variables, there could be noticeable clustering of assigned p's around certain values. While this is hardly objectionable in itself, the chi-square statistic as a goodness of fit test could be quite sensitive to where the K-1 division points on the p range fall with respect to the positions of the cluster peaks.

IV. THE NORTH CAROLINA DATA: Orsagh and Marsden (1984)

A. The Data Base.

In order to test a "rational choice" model of rehabilitation and recidivism, Thomas Orsagh and Mary Ellen Marsden collected information on all men released from prison in North Carolina in the first half of 1980. For their analyses they kept in the study population only those subjects who were under age 50 on January 1, 1980, and had spent at least six months in prison just prior to their release. This yielded a sample of 1,425 individuals, whose criminal justice system contacts and employment history were followed in official records for two years from the date of release.

For the present analysis, recidivism is defined as any arrest that was recorded by the North Carolina Police Information Network.[10] There were 1185 subjects for whom unambiguous information was available on time to failure or censoring and these records were randomly separated into three approximately equal sub-samples. Finally, within each sub-sample only those individuals were retained in the study populations whose records showed no missing data on any of the independent variables included in the analysis.

The independent variables used are defined by Orsagh and Marsden as follows:

Age = Inmate's age on 1 Jan. 1980.

Alc pgmd = 1 if inmate participated in an alcoholics' rehabilitation program prior to exit date. Otherwise = 0.

Alchd = 1 if reported to have been a frequent drinker.

Arr. Rate = arrest rate per year between age 12 and the year of admission on the instant incarceration.

Deterp = regional ratio of property arrests to reported property offenses in 1979. Property offenses are defined as larceny, auto theft, burglary and robbery. The region is that which contains the offender's home county.

Deterv = like deterp but for homicide, rape and assault.

---

[10]This is the state agency responsible for preparation of North Carolina's submissions to the FBI's Uniform Crime Reports. We are informed that PIN records arrests for UCR index crimes.

Dmh pgmd = 1 if inmate ever participated in a drug rehabilitation or mental health program during his instant incarceration. Otherwise = 0.

Drmhd = 1 if the inmate reported a drug problem or had received treatment for a mental health problem. A drug problem is defined as "uses drugs frequently" or "former drug user." A mental health problem is defined as "any history of any mental problem." Definitions are derived from inmate history, compiled by the Department of Correction.

Ed vocn = number of prison educational and vocational training programs the inmate enrolled in while in prison. GED exams are included if taken and passed during the present incarceration. (This variable differs from the Orsagh/Marsden definition by merging the GED variable with the original ed vocn.)

Ed years = the number of years of schooling which the inmate is reported to have had.

Job Skld = 1 if the inmate was employed as a skilled or semi-skilled worker or was a student prior to his instant incarceration. If he was an unskilled worker, unemployed, or reported no occupation, Job Skld = 0.

Marryd = 1 if the inmate was married and living with his spouse at the time of the arrest resulting in the instant incarceration. Otherwise = 0.

Numpty = the total number of property arrest counts relating to offenses committed prior to the instant incarceration as derived from the Police Information Network.

Pracd = 1 if the inmate participated in the post-release component of the Pre-Release and After-Care program. Otherwise = 0.

Race = 0 for "white" and 1 for "nonwhite".

Released = 1 if the inmate was released from prison under supervision on his exit date. Unconditional release -- "maxing out" -- = 0.

Rule brk = the number of reported rule violations per year during the instant incarceration.

Time In = the natural log of the number of years served by the inmate during his instant incarceration, rounded to the nearest quarter of a year.

Total = the total number of counts on all arrests prior to the inmate's instant incarceration. Note that each arrest may have several counts.

Unemploy = the regional unemployment rate for males within the region in which the inmate's county of release was located. Data refer to 1980.

Wrk hisd = 1 if the inmate's reported work history, based on his employment record as coded by the Department of Correction, indicates a stable work record and working regularly at the time of the offense or that he was a student at that time. Any other code = 0.

Wrk Pd = 1 if the inmate participated in one or more prison duty programs or in one or more prison enterprise (industry) programs during his instant incarceration. If he participated in neither, this variable is 0. It has the value 2 if he participated in both types of programs. (This variable is a combination of two separately defined variables in the Orsagh/Marsden data.)

Wrk Reld = 1 if the inmate was ever on work release during his instant incarceration. Otherwise = 0.

Table A gives for the three subsamples the mean values of these variables along with total counts of variables taking only values of 0 or 1 and standard deviations of the remaining variables.

Table A
North Carolina Releases -- Samples A/B/C
N=298/326/279

Variable	Means	Standard Devs.	Totals
Failed	.43/.51/.49		128/166/138
Time to failure or censoring (days)	566/523/530	232/255/248	
(Note: All non-failers were "observed" for 731 days.)			
Age	26/27/26	7.9/7.7/7.1	
Alc Pgm	.09/.12/.13		27/38/37
Alchd	.31/.33/.34		93/108/96
Arr Rate	.23/.24/.24	.16/.17/.16	
Deterp	.20/.20/.20	.047/.048/.044	
Deterv	.75/.74/.73	.20/.21/.20	
Dmh Pgm	.08/.08/.09		24/26/25
Drmhd	.21/.17/.15		63/55/42
Ed vocn + Ged	.55/.51/.52	.57/.53/.56	

(Table A continued)

Ed Yrs	9.8/10.0/9.6	2.4/1.9/2.1	
Job Skld	.44/.46/.46		131/149/129
Marryd	.24/.22/.24		72/72/66
Numpty	2.55/2.50/2.47	2.80/2.48/2.79	
Pracd	.36/.27/.30		107/88/85
Race	.54/.54/.56		161/176/155
Released	.86/.80/.87		255/260/243
Rule Brk	1.18/1.25/1.06	1.9/2.1/1.9	
Time In	.23/.27/.20	.59/.64/.61	
Total	4.45/4.78/4.32	5.0/4.8/4.3	
Unemploy	4.98/4.65/4.72	1.5/1.5/1.5	
Wrk Hisd	.48/.48/.48		143/157/135
Wrk Pd	.80/.84/.77	.57/.57/.63	
Wrk Reld	.45/.53/.48		134/173/135

The three sub-samples obviously differ in certain population characteristics. But there is no a priori reason to suppose that this kind of variability would not also be found in single samples of this size drawn from a hypothetically infinite population of North Carolina Prison releases. Or, to put things in more practical terms, it is assumed that the next 300 or so releases would show a similar variability in their population characteristics.[11]

B. Logit Analysis.

1. Model Construction.

Table B gives the results of the initial logit models constructed separately on each of the samples. The independent variable is arrest for a new offense within two years following release. These are "initial" models in the sense that they are built using all independent variables.

---

[11]Note the assumption here that inter-sample variability is random and that significant characteristics of the underlying population are unchanging in time. In practice this is not likely to hold over long time periods. (Reiss, 1951; Copas and Tarling, 1986) Prudence suggests periodic recalibration of any prediction instrument used in making dispositional decisions.

Table B
Initial Logit Models

	Sample		
	A	B	C
N =	298	326	279
Failures =	128	166	138
Variable	Coefficient Value (t statistic)		
Constant	-2.811	2.196	0.326
	(1.852)	(1.572)	(0.204)
Age	-0.018	-0.047	-0.207
	(0.652)	(1.853)	(0.637)
Alc Pgm	0.121	0.315	-0.020
	(0.250)	(0.770)	(0.048)
Alchd	0.138	-0.034	0.527
	(0.451)	(0.115)	(1.630)
Arr Rate	2.730	0.685	3.716
	(1.660)	(0.493)	(2.287)
Deterp	4.462	-3.505	-0.736
	(1.200)	(0.974)	(0.180)
Deterv	0.977	1.123	0.652
	(1.112)	(1.369)	(0.691)
Dmh Pgm	0.221	-0.187	-0.325
	(0.435)	(0.389)	(0.633)
Drmhd	0.300	0.413	0.414
	(0.876)	(1.118)	(1.024)
Ed vocn	-0.389	0.036	0.076
	(1.384)	(0.137)	(0.259)
Ed Yrs	0.049	-0.130	-0.096
	(0.772)	(1.887)	(1.411)
Job Skld	-0.108	0.066	0.772
	(0.354)	(0.242)	(2.568)
Marryd	-0.221	-0.147	-0.417
	(0.640)	(0.442)	(1.151)

(Table B continued)

Numpty	-0.047	0.095	0.026
	(0.518)	(1.287)	(0.285)
Pracd	-0.587	-0.219	-0.493
	(1.835)	(0.685)	(1.505)
Race	0.199	0.207	0.796
	(0.673)	(0.746)	(2.477)
Released	0.304	-0.310	-0.105
	(0.655)	(0.818)	(0.233)
Rule Brk	0.167	0.316	0.092
	(1.909)	(3.134)	(0.910)
Time In	0.109	0.351	-0.391
	(0.415)	(1.469)	(1.428)
Total	0.158	0.059	-0.010
	(2.001)	(1.198)	(0.150)
Unemploy	-0.046	-0.206	-0.069
	(0.486)	(2.289)	(0.683)
Wrk Hisd	-0.162	-0.096	-0.496
	(0.549)	(0.360)	(1.665)
Wrk Pd	-0.028	0.201	-0.129
	(0.115)	(0.825)	(0.536)
Wrk Reld	-0.411	0.250	-0.412
	(1.391)	(0.859)	(1.288)
Log Likelihood	-173.17	-191.75	-161.43
Accept naive model?[12]			
p =	.000029	.0000022	.000010
df =	23	23	23

[12]The "naive" model assumes all subjects have the same failure probability: the failure rate found in the population on which the model is constructed. The p value is the chance of observing a difference in likelihood values equal to or greater than that between the naive and fitted models, given acceptance of the hypothesis that the model with covariates carries no more recidivism information than the naive model. (Aldrich and Nelson (1984))

Clearly, Table B suggests both consistencies and inconsistencies between the models built on the three different samples. What is not so obvious is the validity of inferences that might be drawn, given the modeling results from just one of these samples.

The analysis proceeds by defining samples A and B to be the construction data and reserving sample C for empirical validation. As explained previously, a decision rule is adopted for successively dropping from the model that variable whose elimination would produce the greatest estimated increase in the sum of log likelihoods of the models built on A and B plus the sum of cross-validation log likelihoods.

Table C shows the values of this sum and the order of elimination of variables.

Table C
Variables successively eliminated in model construction and resulting value of log likelihood sum (U(A,B)).

Variable dropped	Resulting U
(Initial Value)	-784.945
Constant	-778.102
Total	-776.887
Wrk Reld	-773.862
Ed Vocn	-771.980
Wrk Pd	-770.598
Job Skld	-770.479
Deterp	-770.395

At this point, dropping any of the remaining variables would result in a decrease in U.

It would be tedious to reproduce here the results of all 16 models built in the course of this 8 stage process. But a few comments about what happened along the way might not be uninteresting.

In general the effect of dropping the constant term was to decrease values of most of the coefficients estimated on data set A and increase those in model B. No coefficients changed sign at this stage. In particular, the coefficients of "total" became .158 and .039 in models A and B, respectively. From their definitions, we might expect "total" and "numpty" to be correlated in the data. In the stage 2 models (after dropping the constant), the "numpty" coefficients are -.044 and +.090.

On the basis of sign consistency one might then expect that "total" would be kept and "numpty" dropped. The algorithm decided otherwise. After dropping "total", the "numpty" coefficients in the stage 3 model became +.075 and +.107.

The variables that are subsequently dropped through stage 6 have opposite sign coefficients.

In the course of the first 4 stages, model A's coefficient of "deterp" has been positive but decreasing in value. At stage 5 it changes sign. By stage 7 the A and B model coefficients for "deterp" are -.05 and -1.74, respectively. One would expect "deterv" to be the most closely related variable. Its coefficients have not changed much, taking on values of .928 and 1.17 by stage 7. Dropping "deterp" brings these coefficients somewhat closer together at the eighth and final stage with A and B values of .919 and .888.

The results for this final model are given in Table D for samples A and B separately as well as for the solution model estimated on the data of samples A and B combined.

Table D
Logit Solution Model

	Sample		
	A	B	A+B
N =	298	326	624
Failures =	128	166	294

Variable	Coefficient Value (t statistic)		
Age	-0.019	-0.014	-0.016
	(1.103)	(0.846)	(1.351)
Alc Pgm	0.062	0.368	0.260
	(0.132)	(0.916)	(0.886)
Alchd	0.249	0.064	0.141
	(0.869)	(0.222)	(0.710)
Arr Rate	2.974	2.221	2.399
	(2.378)	(2.184)	(3.117)
Deterv	0.919	0.888	0.897
	(1.411)	(1.616)	(2.164)
Dmh Pgm	-0.064	-0.093	-0.131
	(0.132)	(0.199)	(0.395)
Drmhd	0.338	0.344	0.341
	(1.019)	(0.965)	(1.435)
Ed Yrs	-0.046	-0.634	-0.049
	(0.907)	(1.188)	(1.348)

(Table D continued)

Marryd	-0.089	-0.159	-0.110
	(0.270)	(0.505)	(0.494)
Numpty	0.094	0.111	0.101
	(1.318)	(1.604)	(2.079)
Pracd	-0.484	-0.190	-0.326
	(1.593)	(0.623)	(1.571)
Race	0.061	0.179	0.103
	(0.224)	(0.682)	(0.561)
Released	-0.403	-0.160	-0.286
	(1.042)	(0.481)	(1.161)
Rule Brk	0.123	0.324	0.216
	(1.567)	(3.395)	(3.679)
Time In	0.001	0.414	0.249
	(0.003)	(1.922)	(1.605)
Unemploy	-0.140	-0.181	-0.165
	(1.622)	(2.165)	(2.834)
Wrk Hisd	-0.051	-0.026	-0.040
	(0.193)	(0.098)	(0.217)
Log Likelihood	-181.193	-194.524	-380.259
Accept Naive Model?			
p =	.00015	1.8E-7	1.2E-14
df =	16	16	16

The final models on A and B give a more consistent picture of the variables associated with recidivism than do the initial models. In particular, the signs on the coefficients are the same in the two models even if some coefficient values differ considerably.[13]

2. Model Validation.

The next step in the analysis is an attempt to measure how good a job the solution model on the combined data does as a "predictor." For this purpose the coefficients of the solution model were applied to the data of sample C, calculating the recidivism probability that the model associates with each C subject. The overall range of probability values was then divided into 19 non-overlapping and equal length segments. The expected number of failers and the standard deviation of that number, based on the binomial, were then calculated for each of the segments. The results are shown in Table E. (The number 19 has no special significance. The computer program does this

---

[13]What about variables whose coefficients have low t statistics? Would we obtain a more generalizable model by dropping these "non-significant" variables? In simple terms a low t value here means that the log likelihood function is relatively "flat" in certain directions in the neighborhood of the solution values and, therefore, relatively indifferent to small changes in values of low-t coefficients. Under certain distributional assumptions the usual single sample test of "significance" measures how reasonable it would be to assume that this neighborhood of indifference includes the coefficient value zero. One might also test whether it reasonably includes the value 7 or any other arbitrarily chosen number. But even if it does, one would hardly feel justified in substituting that number for the solution value.

The reason the value zero is special lies not so much in the mathematics as in the purpose for which the model is being built. Suppose, for example, the study is an evalution and the coefficient in question measures the efficacy of a very expensive treatment. Policy makers might have second thoughts about any widespread implementation unless it would be unreasonable to interpret the study results as indicating no treatment effect. (The awkward double negative here corresponds to the scientific caution and conservatism typical of the significance levels used in such studies.)

The purpose of the models of this paper is much simpler: to see how much predictive validity can be obtained in assigning failure probabilities on the basis of individual covariates, using a particular algorithm to determine an "optimum" degree of parsimony.

validation in a loop with a variable number of segments. It seemed natural to look at the results for 1 and 10 segment divisions of the range; and, of course, 19 is the next number in this arithmetic series.)

Table E
Validation on Sample C of Logit Solution Model Estimated on A+B

Upper p in segment	Segment n	Failures Observed	Failures Estimated	Std. Dev.
.171	6	1	.908	0.877
.217	19	3	3.802	1.743
.262	19	4	4.617	1.869
.308	21	8	5.966	2.066
.353	27	10	8.851	2.438
.399	27	11	10.159	2.516
.444	20	8	8.412	2.207
.490	21	12	9.788	2.285
.535	26	17	13.395	2.547*
.581	17	8	9.441	2.048
.626	18	13	10.910	2.072*
.672	14	8	9.110	1.783
.717	5	4	3.531	1.018
.762	12	10	8.865	1.522
.808	7	6	5.490	1.088
.853	7	5	5.813	0.992
.899	6	5	5.257	0.806
.944	5	4	4.573	0.625
.990	2	1	1.967	0.179**

* = Estimated and Observed failures differ by more than 1 standard deviation in this segment.
** = Estimated and Observed failures differ by more than 2 standard deviations.

The chi-square test now offers a single statistic for assessing predictive validity. Given the null hypothesis, the test gives the probability of obtaining a chi-square value equal to or greater than the value generated by the observations.[14]

In order for all cells to be well enough populated to apply this test, the first two and the last five segments in the above table

[14]Or in frequentist terms. Suppose we accept the model. Chi-square is a function measuring the discrepancy between the vectors of observed and expected outcomes. The test here then gives the fraction of times in a set of hypothetical, repeated tests that we could anticipate a discrepancy equal to or larger than the one obtained.

were each combined into single segments. The resulting value of chi square is 8.503 and, with 27 degrees of freedom, the associated probability is .9997.

It might be noted in passing that, if the same tests were applied to samples A and B (ignoring the fact that the solution model uses data from these samples), under the null hypothesis the chi-square probabilities are .989 and .995, respectively.

3. Interpretation of Model Results.

If one chooses to accept this model, the simplest question one might ask is how it assesses the relative strengths of the variables in their contribution to failure. As mentioned in an earlier section of this paper, the effects of individual variables can most easily be made apparent by hypothesizing two subjects who are identical except in the measures on one variable. The results for the logit solution of Table D (the A+B model) are given below in Table F. In each case, subject i is identical to j on all except one of the model variables. On that single variable i differs from j by the amount shown under the column heading "d". The difference is always in the direction making i the greater recidivism risk. For continuous variables the "d" values in this illustration were chosen to be close to the standard deviations of these variables found in the data.

Table F
Ratios of Failure Odds of Hypothetical Subjects i and j

Variable	d = z(i) - z(j)	Odds(i)/Odds(j)
Age	-7.5	1.13
Alc Pgm	1	1.30
Alchd	1	1.15
Arr Rate	0.165	1.49
Deterv	0.2	1.20
Dmh Pgm	-1	1.14
Drmhd	1	1.41
Ed Yrs	-2	1.10
Marryd	-1	1.12
Numpty	3	1.35
Pracd	-1	1.39
Race	1	1.11
Released	-1	1.33
Rule Brk	1.9	1.51
Time In	0.62	1.17
Unemploy	-1.5	1.28
Wrk Hisd	-1	1.04

If i and j differ on several variables by the amounts of this

example, the odds ratio is simply the product of the corresponding ratios in the table. For example, if i is 7.5 years younger than j, has a prior arrest rate that is .165 greater, has two years less of schooling and 3 more property offense counts in his record, his odds of being arrested for a new offense within two years are 2.5 times j's odds. And if i happens to differ from j by the amounts d on all variables, the model assesses his failure odds at 38 times j's.

These results do not contain many theoretical surprises. Recidivism odds decrease with age, increase with alcohol or drug abuse histories, increase with extent of prior involvement in criminality, and so forth. Some comment might be offered on a few of the model's conclusions.

With regard to the two alcohol variables, the positive sign on the program variable would reasonably seem to indicate that this is being seen by the model as an indicator of an alcohol problem rather than as a pernicious effect of the program itself. Indeed Orsagh and Marsden comment in their report on the unreliability of the "alchd" variable.

Taken at face value, the opposite signs of the coefficients of "dmh pgm" and "drmhd" would seem to indicate some rehabilitative success for the drug program.

The "deterv" variable is interesting since its positive sign indicates that the comparatively greater risk of apprehension for violent offenses is not perceived as a differential level of threat by these subjects as a group. A simple interpretation of this result, then, is that offense behavior is not much affected but, given an offense, the chance of being arrested is just greater in some jurisdictions than in others.

The "pracd" and "released" coefficients are encouraging since both would say that North Carolina's post-prison policies and programs do indeed have some effect in reducing recidivism over the first two years after release.

"Rule brk" is, perhaps, surprisingly strong in its isolated effect. The reader should keep in mind that this variable is measured as a rate since it would otherwise be hopelessly confounded with "time in." And with regard to "time in," the reader is reminded that this is expressed as the log of the years of the present incarceration term. Thus, the .62 of Table F actually translates into i's term of imprisonment being about twice as long as j's.

Finally, the variable "unemploy" obviously has the wrong sign to lend support to a theory that a general scarcity of jobs would contribute to the likelihood of recidivism. Like "deterv," this is an environmental variable. One might surmise that it stands

as a surrogate for a more complex set of societal conditions. For example, it is not totally implausible to guess that an area unemployment rate might somehow be a rough measure of a jurisdiction's position on a rural-urban scale. But from the data available for this analysis, it is impossible to confirm this or even to figure out which end of such a scale would be more conducive to recidivism.

C. Hazard Models -- The Proportional Hazard Solution.

1. Model Construction.

The hazard models investigated in this paper are based on a function that is log linear in the covariates and has the form of a gamma density in time:

$$\ln h(\underline{z},t) = \underline{z}'(\underline{c} + \underline{a} \ln(t) + \underline{b}t)$$

Here $\underline{z}$ is a vector of covariates characterizing the individual subject; $\underline{a}$, $\underline{b}$, and $\underline{c}$ are vectors of model coefficients; and t is the time to failure or censoring.

If $\underline{a}$ and $\underline{b}$ are restricted in advance so that only the coefficients of the intercept terms are non-zero, one obtains a proportional hazard model -- so called because the ratio of the hazard function for different individuals remains constant in time:

$$\ln h(\underline{z},t) = \underline{z}'\underline{c} + a_1 \ln(t) + b_1 t.$$

Like the logit analysis described above, the investigation of the proportional hazard model started with all variables and proceeded to apply the decision rule using the estimated change in the sum of log likelihoods to eliminate variables inconsistently related to recidivism in the construction data samples A and B. One additional variable, "released," was then dropped simply for reasons of programing convenience[15] and the model reestimated. "Released" had had coefficient values -0.209 and -0.056 with t-statistics -0.748 and -0.264 in samples A and B respectively. The final value of the sum of log likelihoods (the psi function) at the end of this construction phase was -771.631

---

[15]The programs for the analyses reported here were written in Gauss, a PC matrix language created by Lee E. Edlefsen and Samuel D. Jones of Applied Technical Systems. Gauss puts a constraint on the size of matrices that can be handled at any one time. "Programming convenience" simply means that the analyses here were tailored to conform to this constraint rather than adapted to a more general procedure requiring that data be read in in a loop.

-- slightly less than the logit model's -770.395.

The results along with the solution model constructed on the combined data are given in Table G.

Table G
Proportional Hazard Solution Model

	Sample		
	A	B	A+B
N =	298	326	624
Failures =	128	166	294
Variable	Coefficient Value (t statistic)		
Coefficients of term constant in time: c			
Age	-0.011	-0.033	-0.025
	(0.737)	(2.450)	(2.543)
Arr Rate	1.973	0.612	1.013
	(2.342)	(0.919)	(1.971)
Deterv	0.272	0.368	0.319
	(0.606)	(1.050)	(1.172)
Drmhd	0.021	0.401	0.208
	(0.098)	(1.902)	(1.412)
Ed Yrs	-0.010	-0.044	-0.029
	(0.283)	(1.254)	(1.192)
Marryd	-0.113	-0.034	-0.075
	(0.473)	(0.159)	(0.479)
Numpty	0.068	0.065	0.062
	(1.275)	(1.413)	(1.777)
Pracd	-0.490	-0.233	-0.338
	(2.469)	(1.241)	(2.558)
Rule Brk	0.129	0.153	0.142
	(3.095)	(4.480)	(5.621)
Time In	0.008	0.113	0.080
	(0.048)	(0.868)	(0.819)
Total	0.011	0.031	0.025
	(0.386)	(1.152)	(1.409)

(Table G continued)

Unemploy	-0.120	-0.120	-0.127
	(2.580)	(2.086)	(3.119)
time coefficient: b_1			
	-0.939	-0.080	-0.365
	(2.580)	(0.310)	(1.745)
ln time coefficient: a_1			
	0.851	0.144	0.362
	(3.290)	(0.986)	(2.795)
Log Likelihood			
	-261.02	-304.36	-572.21

Comparison with hazard function constant in time and uniform across subjects: Accept null hypothesis?
p(chi-square) = 1.8E-8 4.9E-9 1.5E-9
df = 13 for all 3 samples.

The proportional hazard model solution differs from the logit model by dropping both alcohol variables along with the drug program, race and work history variables. However, it has chosen to retain the variable "total" (number of arrest counts) along with "numpty" (number of property arrest counts). As noted above, "released" was dropped rather arbitrarily by the analyst.

All three models of Table G have a similar time dependence. The hazard function starts at 0, passes through a maximum at $t = -(a_1/b_1)$ and then decays asymptotically to 0. The models on samples A and B do, however, differ somewhat in position and shape. For the A sample, the hazard function passes through a maximum at about 11 months after release and through half its maximum at 2 months and 31 months. For the B sample the maximum occurs at 21 months with the half maxima at .06 and 169 months. The combined sample produces a time dependence closer to sample A's in position but to B's in shape: a maximum at 12 months and half maxima at .7 and 52 months. Essentially, then, the proportional hazard model solution represents a continuously rising risk of recidivism during the course of the first year followed by a quite slow decline in risk after that. It should be recalled that "risk" at any given time here means a probability of recidivism in the near future conditioned on the subject's not having yet recidivated.

2. Model Validation.

Table H gives the results of validation of this model on the data of Sample C.

Table H
Validation on Sample C of Proportional Hazard Solution Model Estimated on A+B

(p = probability of failure within two years after release.)[16]

Upper p in segment	Segment n	Failures Observed	Failures Estimated	Std. Dev.
.175	5	1	0.778	0.810
.221	11	2	2.219	1.330
.267	22	2	5.482	2.029*
.313	21	7	6.173	2.087
.359	22	9	7.413	2.216
.404	29	12	11.027	2.613
.450	36	17	15.359	2.966
.496	26	12	12.229	2.544
.542	24	17	12.420	2.447*
.588	14	9	7.865	1.856
.634	10	9	6.119	1.540*
.679	13	7	8.572	1.708
.725	14	8	9.738	1.721*
.771	13	12	9.745	1.561*
.817	1	1	0.779	0.415
.863	2	1	1.714	0.495*
.908	8	6	7.083	0.900*
.954	4	3	3.747	0.487*
1.00	4	3	3.944	0.233**

* = Estimated and Observed failures differ by more than 1 standard deviation in this segment.
** = Estimated and Observed differ by more than 2 standard deviations.

Collapsing the first two and the last five segments into single elements, one obtains a chi-square of 22.17 with 27 degrees of freedom and probability of .729 under the assumption of the null hypothesis. Again, if this model were applied to samples A and B as a test of goodness of fit, the respective chi square probabilities are .999 and .925.

---

[16]The test based on probability of failure within a fixed two year period allows straightforward comparisons between logit and hazard models. Results specifically aimed at examining the validity of the hazard model's time dependence are given in section E. below.

3. Interpretation of Model Results.

Single variable effects in this model can again be isolated by considering the hazard ratios for two individuals who differ only in their values on one of the modeled variables. For the proportional hazard model these ratios remain constant in time. They are, of course, to be interpreted as ratios of conditional probabilities, valid for all time t under the assumption that both i and j have survived to t.

The results are given in Table I.

Table I
Ratios of Conditional Failure Probabilities of Subjects i and j
Proportional Hazard Model

Variable	d = z(i) - z(j)	Failure Probability Ratio p(i)/p(j)
Age	-7.5	1.21
Arr Rate	0.165	1.18
Deterv	0.2	1.07
Drmhd	1	1.23
Ed Yrs	-2	1.06
Marryd	-1	1.08
Numpty	3	1.20
Pracd	-1	1.40
Rule Brk	1.9	1.31
Time In	0.62	1.05
Total	4	1.11
Unemploy	-1.5	1.21

These results are similar to the corresponding logit model odds ratios reported in Table F. If hypothetical subject i happens to be so unfortunate as to differ from j by the amounts of Table I on all variables, his conditional failure probability would be 6.6 times greater.

D. The "Full" Hazard Model Solution.

1. Model Construction.

As mentioned above, the proportional form of the hazard model is obtained by imposing a restriction on the coefficients of the time dependent terms in the hazard function. This is quite a strong assumption to make about the recidivism process. It implies that all subjects, no matter how different they may be in terms of their values on the model's independent variables, will pass through the point of maximum risk at the same time. The spread of the risk function (as determined by the times at which

the risk passes through its half maximum value) will also be identical. Individual covariates can determine only the relative heights of the risk curves.

We now want to investigate a hazard model that allows the data on the individual subjects more freedom to determine the time dependence of their recidivism risk:

$$\ln h(\underline{z},t) = \underline{z}'(\underline{c} + \underline{a} \ln(t) + \underline{b}t).$$

If an intercept term and all 24 independent variables used in the logit and proportional hazard model estimations were included in each of the 3 coefficient vectors, a total of 75 model parameters would have to be calculated. The data samples A and B did not seem large enough to produce that many analytically stable results. The construction phase, therefore, begins with the variables retained in the proportional hazard solution but restores the variable "released" and tests for the effect of an alcohol problem by including "Alc Pgm."

At the end of the construction run, "released" was again arbitrarily dropped for programing convenience and the coefficients reestimated. This variable had been retained in the model only in the linear term in time with coefficients -.043 and -.017 and standard deviations .432 and .468 for samples A and B.

Finally, three more terms were dropped from the model because of very small coefficient values: "rule brk" in the linear term with a coefficient of -.0039 and "rule brk" and "total" in the log term with coefficients -.00032 and -.0038, respectively. The results estimated on this final set of variables are shown in Table J.

Table J
Full Hazard Solution Model

	Sample		
	A	B	A+B
N =	298	326	624
Failers =	128	166	294
	Coefficient Value (t statistic)		
Time Independent Term: $\underline{c}$			
Constant	-1.573	-1.111	-1.398
	(2.892)	(2.582)	(4.278)
Arr Rate	3.333	2.439	2.489
	(2.668)	(2.223)	(3.206)

(Table J continued)

Marryd	-0.113	-0.115	-0.103
	(0.464)	(0.544)	(0.662)
Numpty	0.371	0.115	0.281
	(1.577)	(0.693)	(2.321)
Pracd	-0.387	-0.210	-0.296
	(1.956)	(1.107)	(2.219)
Rule Brk	0.296	0.093	0.143
	(1.803)	(0.876)	(5.642)
Time In	0.347	0.432	0.489
	(0.428)	(0.896)	(1.191)
Total	-0.120	-0.039	-0.066
	(0.984)	(0.414)	(2.103)
Unemploy	-0.066	-0.105	-0.093
	(1.014)	(1.792)	(2.182)
Linear Term in Time: b			
Age	-0.023	-0.015	-0.015
	(1.730)	(1.351)	(1.865)
Alc Pgm	0.108	0.149	0.123
	(0.383)	(0.587)	(0.662)
Arr Rate	-2.035	-1.432	-1.287
	(1.477)	(1.243)	(1.573)
Deterv	0.908	0.418	0.614
	(2.218)	(1.301)	(2.502)
Drmhd	0.196	0.014	0.123
	(1.017)	(0.059)	(0.867)
Numpty	-0.304	-0.065	-0.162
	(1.458)	(0.439)	(1.872)
Rule Brk	-0.191	0.079	omitted
	(1.219)	(0.770)	
Time In	-0.362	-0.202	-0.288
	(0.370)	(0.466)	(0.793)
Total	0.147	0.063	0.093
	(1.268)	(0.739)	(2.784)

(Table J continued)

Log Term in Time: a

Deterv	0.771	0.155	0.287
	(1.978)	(0.754)	(1.602)
Drmhd	0.327	-0.414	-0.198
	(1.067)	(2.849)	(1.529)
Numpty	0.075	8.4E-5	0.027
	(0.512)	(0.001)	(0.820)
Rule Brk	0.042	-0.010	omitted
	(0.557)	(0.231)	
Time In	0.154	0.369	0.361
	(0.324)	(1.574)	(1.689)
Total	-0.025	0.025	omitted
	(0.336)	(0.526)	
Log Likelihood	-248.860	-300.870	-563.063

Comparison with hazard function constant in time and uniform across subjects: Accept null hypothesis?

p(chi square) =	3.7E-9	3.4E-7	9.2E-20
df =	23	23	20

At the end of the construction phase (but before omission of the three terms with small coefficients) the log likelihood sum (U(A,B)) was -765.598 -- an algebraically slightly greater value than that obtained with either the logit or proportional hazard models.

2. Model Validation.

The results of the validation of the A+B model on sample C data are given in Table K. Again, as in the validation results for the proportional hazard model, the results here are given in terms of modeled probabilities of failure within two years after release from prison.

Table K
Validation on Sample C of Full Hazard Solution Model
Estimated on A+B

Upper p in segment	Segment n	Failures Observed	Failures Estimated	Std. Dev.
.187	3	1	0.475	0.632
.232	10	1	2.141	1.297
.277	22	2	5.617	2.045*
.322	29	14	8.594	2.458**
.368	24	7	8.335	2.332
.413	32	14	12.622	2.764
.458	42	19	18.297	3.212
.503	23	14	11.002	2.395*
.548	18	10	9.441	2.118
.593	17	12	9.701	2.040*
.639	9	7	5.460	1.465*
.684	14	9	9.231	1.773
.729	10	8	7.015	1.447
.774	6	5	4.513	1.057
.819	0	-	--	--
.864	7	6	5.919	0.955
.910	5	3	4.400	0.726*
.955	4	4	3.739	0.494
1.00	4	2	3.930	0.261**

* = Estimated and Observed failures differ by more than 1 standard deviation in this segment.
** = Estimated and Observed failures differ by more than 2 standard deviations.

Taking the first three and the last seven segments as single divisions of the probability range, one obtains a chi square of 12.8. Under the null hypothesis the chi square probability with 21 degrees of freedom is .92. By this test the validation fit is not as good as the .99 obtained with the logit model but considerably better than the proportional hazard's .73. The proportionality assumption for the recidivism process does indeed seem dubious.

3. Interpretation of Model Results.

Again individual variable effects will be investigated by considering the hazard function ratios for two individuals who are identical on all variables but one. The problem is more complicated than in the proportional hazard case since the ratios may be changing over time. Each variable of the solution model will, therefore, be discussed separately. The ratios that are given as examples in the discussion should be considered as short

term results, valid, say, for the ensuing month. And again the reader is reminded that these are ratios of conditional probabilities and assume that both subjects have survived to the time in question.

Age: The hazard ratio of the 7.5 years younger to the older subject increases exponentially in time from an initial value of 1. At t = 1 year, the ratio is 1.12; at t = 2 years, it increases to 1.26.

Alc Pgm: For a subject with an alcohol problem as evidenced by his participation in an alcohol program during the present incarceration, the hazard ratio again increases exponentially from an initial value of 1 when compared with the subject without such evidence of a problem. At one year and two years following release, the hazard ratios are 1.13 and 1.28 respectively.

Arr Rate: Initially, the individual with a prior arrest rate greater by .165 has a failure probability 1.5 times as great as his comparison subject. This decreases exponentially so that at the end of the first year their risk ratio is 1.2 and after two years the risks are virtually identical.

Deterv: With a difference of 0.2 in the clearance rates for violent crimes in their respective jurisdictions, the log of the hazard ratio of the two subjects as a function of time is

$$\ln h(i)/h(j) = .2(.614*t + .287*\ln(t)).$$

The ratios after one and two years are 1.1 and 1.3. Because of the positive coefficient of the ln(t) term, the hazard ratio starts at zero. At two months it is .92 and passes through the point of equal risk at about 5 months. Without putting too much credence in all this, it is still not uninteresting that, taken at face value, this model purports to detect a short-lived and rather weak deterrence effect. Any differential inhibitions of behavior wear off rather quickly, however, and a system effect takes over in which the probability of arrest given a crime is simply greater in some jurisdictions than in others.

Drmhd: Here the ratio starts high, passes through a minimum at about 19 months and then begins to increase again. The person recorded as having a drug or mental health problem is the subject at greater risk throughout. The functional form of the ratio is

$$\ln h(i)/h(j) = .123*t - .198*\ln(t)$$

The values of this ratio are given below for selected times.

t (months)	h(i)/h(j)
2	1.455
12	1.131
19	1.109
24	1.115

A result like this might be "explained" by a gradual recidivism to drugs followed by a recidivism to crime. However, a rather more credible explanation is that the U-shape is simply a consequence of the limitations of the hazard's functional form: it cannot represent a function that is monotonically decreasing to an asymptote other than zero. It looks suspiciously like the model may be trying its best to do just that.

Marryd: Here the hazard ratio remains constant in time with the single subject's risk being about 1.1 times that of his married counterpart -- essentially the same as the proportional hazard model result.

Numpty: This is the number of all property offense arrest counts in the subject's record. Orsagh and Marsden introduced this variable under the hypothesis that the offender dedicated to property crime has a higher probability of recidivism than does his less specialized counterpart. It is to be noted that this model retains both "numpty" and "total" (the number of all prior arrest counts). Therefore, in comparing here two subjects who are identical on all variables other than "numpty", we are comparing individuals with the same "total", one of whom happens to have more property counts in his record than does the other. If a "numpty" difference of 3 between subjects i and j is assumed, the time dependence of the hazard ratio is given by:

$$\ln h(i)/h(j) = 3*(.281 - .162*t + .027*\ln(t)).$$

This is an ordinary gamma density curve, starting at 0, passing through a maximum at t = 2 months and declining asymptotically to 0. The spread about the maximum is very wide. The curve passes through its half maximum values at 1.3 minutes and at 5 years. Clearly, the model is doing its best to represent a hazard ratio curve that starts out finite and with a very small initial slope, and then decreases over the course of time.

With this interpretation imposed on the results, the model says that in the early months subject i's risk is a substantial 1.85 times j's but by month 22, if i and j both survive that long, their risks are virtually identical.

Pracd: The model gives a hazard ratio that remains constant in time. The subject who did not participate in the post release program has a conditional probability of recidivism about 1.3 times that of the program participant.

Rule Break: As noted above, the time dependent effect of this variable on the hazard function was so weak that it was decided to consider its contribution as constant in time. With a difference of 1.9 in the rate of rule violations over the course of the present incarceration, the hazard function for the individual with the higher rate of violations is 1.3 times that of his more conforming counterpart. The same ratio was obtained with the proportional hazard model.

Time In: The model gives this variable a gamma density shape in time. With a difference of .62 in the log of the present term of incarceration (i's term being almost double j's), the hazard ratio becomes:

$$\ln h(i)/h(j) = .62*(.489 - .288*t + .361*\ln(t)).$$

This curve passes through its maximum at about 15 months and through half its maximum at 1 month and at 5 1/2 years. For about the first four months after release, the model credits the individual who served the longer time with a lower failure probability. At 15 months the ratio takes its maximum value of 1.14, decreasing after that so that i and j would have about the same risk (given survival) 3 years after release -- which is, of course, well after the termination of observations in this data.

The effect is clearly not a very strong one, especially when one considers the strength of the "treatment" in this example in its relation to the average experience of the population -- terms of about 15 months. It seems unlikely that subjects i and j could be essentially identical in all other significant respects and still have spent such different lengths of time in prison during their instant incarceration.

Total: With a difference of 4 in the total number of counts in their arrest histories (but the same number of property crime arrest counts), the log of the hazard ratio over time is given by the model as:

$$\ln h(i)/h(j) = 4*(-.066 + .093*t).$$

This ratio starts at a value of .77 and then increases exponentially, passing through the point of equal risk at about 8.5 months. At 12 and 24 months, the ratios are 1.1 and 1.6 respectively. It might be noted that both of the model coefficients here have quite solid t-statistics. Further, the difference of 4 prior arrest counts for non-property offenses is just about equal to the average number of total arrest counts across the population. Hypothetical subject i would seem to be a rather tougher case than j and the exponential increase in their recidivism risk ratio over time (always given survival of both to the time in question) is, perhaps, believable enough. However,

the reality and the interpretation of the inversion of their risk ratio during the first three quarters of a year following release are puzzling.

Unemploy: The model says that the effect of the local unemployment rate on recidivism risk is constant in time and that, with a difference of 1.5 in this rate, the individual in the economically more favored jurisdiction has a conditional probability of recidivism that is about 1.15 times that of his counterpart in the area of greater unemployment. This same result was obtained with the proportional hazard model.

E. Speculations on Policy Use of Hazard Models.

An indication might be given of how such a model could be used for establishment of a policy of differential allocation of resources to post-release supervision. This is simply an illustration and by no means suggests that the authors consider this particular model an adequate basis for actual development of such a policy.

The hazard function at any time t gives a score that expresses the model's estimate of the near term risk of failure of those subjects who are still considered at risk. We imagine, then, a policy that at any time t sorts the subjects into four classes:

1. Those who have already failed, say through arrest for a new crime, but are still under supervised release.

2. Those whose hazard values lie above some policy-determined upper cut point.

3. Those subjects whose hazard values lie below this point but above some policy-determined lower cut point.

4. Those with hazard values below this lower cut point.

The policy, obviously, would consist in a distribution of resources so that the intensity of supervision decreases from class 1 through class 4.

Under the assumption that the probability of recidivism in fact decreases with the increased intensity of supervision, is there any evidence in the analysis of the North Carolina data to suggest that such a policy might work? Do the people who fail tend in general to have a higher hazard score at the time of failure than do the successes at time of censoring?

One bit of evidence suggesting that this is so is provided by the mean values of the hazard at time of failure for those who failed and at time of censoring for those who did not. Using the model

built on sample A+B but applying it to the whole population of 903 subjects, we obtain:

Mean h of failures at time of failure:	.536
Standard deviation:	.615
Mean h of successes at time of censoring (731 days for all):	.324
Standard deviation:	.587

The t-test for difference in means associates a probability of 6.4E-8 with the null hypothesis.

A rather more detailed picture of how such a policy might work in practice is given in Table L below.[17] We are assuming a scenario in which individuals are reclassified according to their hazard scores at four week intervals. Definitions of the column entries are:

Col. 1. Days since release at the beginning of the interval.

Col. 2. Number of subjects still at risk at the beginning of the interval.

Col. 3. Observed failures within the ensuing four weeks.

Col. 4. Expected failures within the ensuing four weeks along with the standard deviation. As in previous tables, the expected number of failures and the standard deviation are based on the binomial distribution. Using the hazard function assigned to each individual who is still at risk, we calculate the probability of failure within the next 28 days. (The mathematics of this calculation is described in Note 1 at the end of this section.)

Col. 5. Mean probability of failure and standard deviation for individuals who failed during this interval. (In the table entries these are multiplied by 100.)

Col. 6. Mean probability of failure and standard deviation for individuals who survived this interval.

Col. 7. Probability associated with t-test of null hypothesis for differences in the means reported in columns 5 and 6.

---

[17] In Table L the entire study population was used to give a relatively fine-grained picture of how the modeled probabilities are changing in time. The Table in the appendix to this section gives similar information for the results of the (A+B) model applied to the data of the validation sample (C). But there the time period is divided into 8 intervals of 91 days each.

Table L
4 Week Interval Comparisons of Observed and Expected Failures
Among Popultation Still at Risk
(Data Sample A+B+C. Model built on A+B.)

t1 days	N(t1)	Obs. fail	Exp. fail. (s.d.)	Av. P(fail) (s.d.) x100	Av. P(surv.) (s.d.) x100	p(t)
1	903	9	14.3* (3.7)	3.43 (2.27)	1.57 (2.15)	.007
29	894	19	19.1 (4.3)	7.68 (9.22)	2.02 (2.16)	.004
57	875	29	20.3* (4.4)	5.94 (6.25))	2.19 (2.15)	.0006
85	846	30	20.2** (4.4)	3.03 (1.82)	2.37 (2.36)	.026
113	816	25	20.5* (4.4)	3.90 (3.70)	2.46 (2.44)	.027
141	791	8	20.3** (4.4)	6.17 (4.63)	2.53 (2.49)	.013
169	783	17	20.4 (4.4)	3.11 (3.31)	2.60 (2.53)	.26
197	766	24	20.3 (4.4)	3.34 (3.31)	2.63 (2.53)	.13
225	742	17	19.8 (4.3)	5.76 (6.21)	2.67 (2.40)	.018
253	725	19	19.1 (4.3)	2.95 (1.65)	2.63 (2.46)	.20
281	706	19	18.7 (4.2)	4.44 (2.41)	2.60 (2.42)	.0005
309	687	21	18.0 (4.1)	3.37 (2.53)	2.60 (2.52)	.082
337	666	16	17.4 (4.1)	2.84 (2.60)	2.60 (2.57)	.36
365	650	19	17.0 (4.0)	4.87 (6.79)	2.55 (2.37)	.069

(Table L continued)

393	631	18	16.1 (3.9)	2.65 (1.53)	2.56 (2.42)	.41
421	613	8	15.7* (3.9)	3.74 (1.73)	2.55 (2.45)	.027
449	605	15	15.5 (3.8)	3.58 (2.41)	2.53 (2.49)	.048
477	590	13	14.9 (3.8)	5.49 (6.90)	2.47 (2.32)	.057
505	577	16	14.3 (3.7)	2.69 (1.02)	2.47 (2.39)	.21
533	561	9	13.8* (3.6)	3.10 (1.07)	2.46 (2.46)	.042
561	552	13	13.6 (3.6)	2.88 (1.50)	2.45 (2.54)	.16
589	539	7	13.2* (3.5)	3.48 (1.22)	2.44 (2.61)	.015
617	532	16	13.0 (3.5)	2.75 (1.40)	2.44 (2.71)	.20
645	516	15	12.6 (3.4)	3.70 (2.28)	2.41 (2.79)	.016
673	501	15	12.1 (3.4)	2.58 (0.71)	2.40 (2.90)	.22
701	486	14	11.7 (3.3)	3.46 (2.95)	2.38 (2.97)	.088

* = Expected and Observed Failures differ by more than one standard deviation.
** = Expected and Observed Failures differ by more than two standard deviations.

There is no single statistic (known to the authors, at least) to measure the overall goodness of fit between the sequences of

observed and expected failures.[18] The model doesn't seem to be doing too bad a job, considering that failure in any given four week period is a relatively rare event. Some further insight might be obtained by comparison with the results obtained from a "naive" model in which it is assumed that the number of failures in each interval is a constant fraction of the population surviving up to that time. Fitting this naive model at the two end points (903 subjects initially, 472 survivors after 26 intervals), we obtain .0246 as an estimate for this fraction. (See Note 2 at the end of this section.)

Such a model must obviously produce a monotonically decreasing sequence of expected failures. The values obtained decrease from 22.3 in the first interval to 12.0 in the last. In a comparison with the hazard function results of Table L, the naive model's expected numbers of failures are greater during the first 4 intervals, fall somewhat below the hazard model's estimates through interval 18, and are virtually identical in both models thereafter. Except for the naive model's very weak start, there would be little reason to prefer one model over the other simply in terms of its ability to explain the sequence of failures observed. Indeed, as shown in Note 2, this might have been anticipated if most of the individual hazard functions are not changing too rapidly in time and the overall observation time is not too long.

What is of interest for the triage policy application is not so much the model's fit to the sequence of observed numbers of failures but the question of whether the model does a sensible job of assigning differential levels of risk to subjects in the population at risk. The naive model by definition does not attempt to do so. Some evidence in support of the fitted hazard model's ability to differentiate between individuals is contained in the results shown in the last three columns of Table L.

In each of the 26 intervals the means of the conditional failure probabilities of those who failed are greater than the means of the survivors. Because of this consistency, the differences must be regarded as systematic even though the t-statistics for differences between means indicate that in only half of the 26 intervals is this difference significant at the 95% confidence level. Furthermore, the mean p of the failures remains throughout greater than the .0246 estimated under the naive model's assumption of recidivism as a purely random process.

For the survivors, the mean conditional failure probability

---

[18]Successive intervals are not independent since each subject remains in the population through the interval in which he fails. Successes, of course, are contained in all intervals.

increases monotonically for the first 9 intervals and then begins to decrease quite slowly.

In comparing the sequences of average hazard scores of failers and survivors, it is important to keep in mind that under this model two things should be going on. First, the surviving population should gradually be depleted of some of its higher scoring members. And simultaneously, the individual hazard scores are changing in time. The mathematical form assumed for the model allows individual hazards to take one of four shapes as functions of time. Table M shows how these four functional forms were distributed by this model among subjects who eventually failed and those who did not.

Table M
Distribution of Subjects Among Different Forms Of Hazard Function

	Failers	Successes
Monotonically Increasing	81	114
Monotonically Decreasing	14	11
U-Shaped	21	23
"Gamma density"	316	323

"Gamma density" means a hazard rate that starts at 0, passes through a maximum and then declines asymptotically to 0. For 61 of the failers and 89 of the successes with this type hazard function, the maximum is not reached until some time t greater than the two years over which follow-up data was collected. So, according to the model, 203 (= 114 + 89) of the study's 471 successes had a risk score that was increasing throughout the two year follow-up period. This doesn't imply that these models can see beyond the data on which they are built. Prediction in that sense is not being tested in this paper. It does suggest, however, that a longer follow-up time might well be warranted in analyzing recidivism among a population similar to the North Carolina releasees studied here.

Note 1: On the Probability Calculations in Table L.

The problem is to determine the probability of failure during some finite time interval $[t_1,t_2]$ conditioned on survival to t_1. We define $H(t_1,t_2)$ to be this conditional probability. With $f(t)$ denoting the unconditioned probability density for failure and $S(t)$ the unconditioned probability of survival to time t,

$$H(t_1,t_2) = \frac{\int_{t_1}^{t_2} f(t)\ dt}{S(t_1)}.$$

Since

$$f(t) = -\ dS/dt,$$

this becomes:

$$H(t_1,t_2) = 1 - \frac{S(t_2)}{S(t_1)}\ .$$

Note 2: On the "Naive" Model.

Let n_k be the surviving population expected at the beginning of the k^{th} interval and f_k the estimated number failing during that interval. We define the naive model by assuming that

$$f_k = \alpha n_k$$

where alpha is a constant. Essentially, this assumption regards recidivism as a process of random sampling from the survivor population with a fixed sampling fraction. From this definition it follows that

$$n_{k+1} = n_k - f_k = (1 - \alpha)n_k$$

or

$$n_{k+1} = (1 - \alpha)^k\ n_1.$$

Here we determine alpha simply by making the model fit the initial and final surviving populations:

$$\ln(1-\alpha) = \frac{\ln\left(\frac{n_{27}}{n_1}\right)}{26}.$$

With $n_1 = 903$ and $n_{27} = 472$, alpha equals .0246.

To examine the relation this has to the fitted hazard model, define $H_i(k)$ to be the probability of failure of subject i during interval k, conditioned on his surviving the first k-1 intervals. Using the result of Note 1 and summing over all subjects still at

risk at the beginning of the k^{th} interval, we obtain

$$\sum_{l \in n_k} H_l(k) = n_k \left(1 - \frac{\sum_{l \in n_k} \frac{S_l(t_2)}{S_l(t_1)}}{n_k} \right)$$

where t_1 and t_2 are the beginning and end points of the interval.

Using the definition of $S_i(t)$ and the mean value theorem, we can write without loss of generality

$$\sum_{l \in n_k} H_l(k) = n_k \left(1 - \frac{\sum_{l \in n_k} \exp(-h_l(k)^* T)}{n_k} \right)$$

Here $T = (t_2 - t_1)$ = constant interval length, and $h_i(k)^*$ is some (unknown) value taken on by subject i's hazard function in the course of the k^{th} interval. The result holds for any hazard function that is continuous over the time interval.

The sum on the left is, of course, the expected number of failures in the k^{th} interval under the fitted model. The term in parentheses on the right is the average failure probability taken over the population at risk during that interval. If most of the individual hazard functions are not changing too rapidly in time around interval k, this average might to a good approximation be considered as constant for a set of adjacent intervals. Provided the number of intervals in this set is not too large so that the higher risk population is substantially depleted through failure, the expected numbers of failures in each interval of the set reduces in this approximation to the form of the equation on which the naive model is based. It should be noted, however, that nothing in this approximation requires the strong assumption that individual subjects have the same risk -- an assumption that is necessarily implied by the naive model.

APPENDIX

13 Week Interval Comparisons of Observed and Expected Failures Among Population Still at Risk

Data Sample = C
Model = Full Hazard Solution on Data Samples A+B

t1 days	N(t1)	Obs. fail.	Exp. fail. (s.d.)	Av. P(fail) (s.d.) x100	Av. P(surv) (s.d.) x100	p(t)
1	279	21	18.4 (4.0)	14.1 (12.0)	6.0 (5.8)	.001
92	258	18	20.7 (4.2)	8.4 (3.1)	8.0 (7.7)	.35
183	240	17	20.7 (4.2)	11.3 (9.2)	8.4 (7.8)	.11
274	223	32	19.2** (4.0)	11.8 (8.4)	8.1 (7.3)	.01
365	191	10	15.6* (3.7)	8.7 (4.2)	8.1 (7.3)	.34
456	181	13	14.8 (3.6)	10.6 (6.1)	8.0 (7.5)	.07
547	168	13	13.4 (3.4)	8.5 (3.3)	8.0 (7.9)	.31
638	155	14	12.4 (3.2)	11.3 (8.7)	8.0 (8.0)	.06

* = Expected and Observed Failures differ by more than one standard deviation.
** = Expected and Observed Failures differ by more than two standard deviations.

V. CALIFORNIA YOUTH AUTHORITY DATA: Haapanen and Jesness (1982)

A. The Data Base.

The Orsagh and Marsden North Carolina data used in the previous section allowed models to be built that were rich in covariates but limited to a two year follow-up observation period. In this section the models are built using a very small set of explanatory variables but the data contain recidivism information for a period of eight to ten years following parole.

These data were initially analyzed by the California Department of the Youth Authority to assess the feasibility of the early identification of chronic adult offenders.[19] Male youths released from California Youth Authority institutions (Preston School of Industry, Northern California Youth Center and Fricott Ranch) during the 1960s and early 1970s constituted the sampling base.[20] During their stay with the Youth Authority, psychological, behaviorial and demographic information was collected. Follow-up arrest data were obtained on these individuals in the late 1970s and 1980. These follow-up data were obtained primarily from official arrest records of the California Bureau of Criminal Investigation with supplemental data received from the Federal Bureau of Investigation and the California Bureau of Vital Statistics.[21]

The present analyses were conducted using only the data from the Preston School of Industry. From the original sample of 1715 subjects, 1699 cases contained information on relevant covariates

---

[19]The data were not originally collected for this purpose. Most background and behavioral data were collected by the institution at the time of incarceration for purposes of assessing program effectiveness.

[20]It should be noted that the California Youth Authority is essentially the "last stop" for delinquent youths, often after previous interventions by local and county authorities. Consequently, these youths represent the more serious or habitual offenders and thus are not representative of the entire delinquent population.

[21]A detailed description of the study can be found in Haapanen and Jessness's final report: Early Identification of the Chronic Offender (1982).

and subsequently served as our final sample.[22]

The common event initiating the time at risk is the date of parole from the Preston School with time to failure or censoring reported in days. For the logit model and the fixed-time validation of the hazard models it was decided to use an observation period of 8 years following release. Inferences to be drawn from such models require that all "successes" have the same time at risk. The exclusion of cases censored earlier than 8 years after parole resulted in a final sample size of 1633 for all fixed-time analyses.[23] It should be noted that individuals whose first arrest occurs after 8 years are by definition considered "successes" in these analyses. For the hazard models all 1699 cases were available for use in parameter estimations.

Recidivism has been defined in numerous ways in the literature. For example, Waldo and Chiricos (1977) offer eighteen different operationalizations. The data allowed four operationalizations, representing varying degrees of restrictiveness in the behaviors deemed failures:

1. Any arrest
2. Arrest with a subsequent conviction for that incident, although not necessarily for the arresting offense.
3. Arrest for a felony offense regardless of the subsequent disposition of the case.
4. Arrest for a felony with subsequent conviction, whether or not the conviction was for the arresting offense.

The analytic models of this section were built using the third of these operationalizations. Appendix A lists those offenses used in the construction of this variable.

The explanatory variables used in this study represent a parsimonious set of variables that have been identified as associated with recidivism (Pritchard, 1979): race, commitment type, age at parole, number of prior arrests and age at first arrest.

---

[22]After one year of good behavior individuals may petition the court to have juvenile records sealed. Through inadvertence on our part 27 such cases are included among the 1699 used in the analyses of this section.

[23]For example, in the full sample 110 youths are reported to have died during the follow-up period. For these subjects the mean time to death following parole was 5 years 4 months. The fixed time analyses include any of these individuals who are also reported to have been arrested for a felony within 8 years but exclude those who died within that period and whose records give no indication of a prior failure.

Race is a "dummy" variable coded 1 if the youth was white, 0 if nonwhite.

Commitment type is a four category variable representing the offense for which the individual was arrested and sentenced to the Youth Authority. Values of 1 were coded for violent crimes, 2 for violent-economic offenses, 3 for property and 4 for minor offenses.[24]

Age at parole with a range of 13 to 22 years is self-explanatory.

Number of prior arrests ranges from 0 through 15. Dr. Haapanen points out that these are arrests contained in the records of the California Bureau of Criminal Investigation. To that extent they may be regarded as arrests for serious offenses. But what constitutes a serious offense may vary considerably from one reporting jurisdiction to another.

Age at first arrest has a range of 9 to 20 years. Again this is the first arrest reported in the State records.

Table N gives summary statistics both for the reduced sample of 1633 subjects used for the 8 year follow-up analyses and for the full sample of 1699 subjects used in the hazard model estimations. Failure here is a felony arrest (definition 3).

---

[24]See Appendix C to this section.

1) Violent offenses are offenses 1 through 7.
2) Violent-economic are offenses 8 through 12.
3) Property are offenses 13, 15-17, 26 and 34.
4) Minor offenses are all others.

In the case of multiple charges only the most serious charge in each incident was coded in the original data collection.

Table N
Means and standard deviations
Full and Reduced samples

	Reduced Sample (8 follow-up)	Full Sample
N	1633	1699
Failures	1346	1377
Variable	x (sd)	x (sd)
failure 3	0.824	0.810
race	0.488	0.488
commitment type	3.090 (1.090)	3.091 (1.091)
age at parole	17.564 (1.104)	17.566 (1.104)
prior arrests	1.388 (1.754)	1.347 (1.740)
age first arrest	15.620 (1.698)	15.640 (1.710)

The racial composition of the study population is approximately 50% white and 50% nonwhite with most youths committed to the institution for property offenses. On average these youths had one recorded arrest prior to their present commitment and experienced their first recorded arrest in their fifteenth year. In addition, on average these youths were between seventeen and eighteen years of age when released from Preston.

Appendix B contains summary information on the three other operationalizations of recidivism not employed in the present analyses. This information is included only to allow for a more complete understanding of the Preston sample. Appendix C lists all the offenses for which any subjects were arrested (failure 1). They include crimes against persons and property, sex offenses, auto and vehicle violations, liquor violations and drug and status offenses. As Appendix B indicates, approximately 93% of the Preston sample were arrested for one or more of these offenses subsequent to their parole.

The second operationalization (failure 2) requires a conviction for the definition of failure. (See Appendix D for dispositions

employed in the construction of this variable). Ninety percent (1532) of the individuals released from Preston were later convicted of some violation.

Finally, failure 4 considers as "failures" only those individuals who were subsequently arrested for a felony and were convicted, whether or not the conviction was for the arresting offense. Seventy-two percent of the Preston sample met this criterion of recidivism.

Table O reports separately for failures (felony arrests) and successes the means and standard deviations for all variables used in the analysis. For those individuals who subsequently recidivated, the average time to arrest for a felony was approximately two years (712 days) with a range of one week to thirteen years.

Table O

Means and Standard deviations for successes and failures

	Reduced Sample (8 follow-up)	Full Sample
N =	1633	1699
Failures =	1346	1377
Successes =	287	322
Variable	x (sd)	x (sd)
race		
failures	0.458	0.460
successes	0.631*	0.609*
commitment type		
failures	3.091 (1.088)	3.092 (1.087)
successes	3.087 (1.098)	3.090 (1.111)
age at parole		
failures	17.536 (1.092)	17.535 (1.097)
successes	17.693 (1.151)*	17.696 (1.125)*
prior arrests		
failures	1.534 (1.786)	1.513 (1.778)
successes	0.704 (1.409)*	0.637 (1.354)*
age first arrest		
failures	15.481 (1.669)	15.492 (1.669)
successes	16.275 (1.684)*	16.276 (1.740)*

* significantly different: p less than .05
In both samples "successes" were significantly different from failures in racial composition. Nonwhites were more likely to be

failures than were white youths (t=5.37 reduced sample; 4.85 sample full). The number of recorded prior arrests for successes and failures also differed significantly in both samples. Successes had fewer prior arrests than those who recidivated (t=8.62 and 9.81). Finally, those individuals who did not recidivate were older at the time of their recorded arrest than were recidivists (t=7.31 and 7.53).

No significant differences were observed in regard to the committing offense. And the differences in means of age at parole, while statistically significant, are too small to be meaningful. (Appendix E presents means and standard deviations for successes and failures for the other three recidivism operationalizations).

A life table represents a convenient, non-parametric way of examining the distribution of failures over time. One may expect that with recidivism data failure will tend to occur in the early years following parole, with the risk of being arrested inversely related to the length of survival.

Table P presents three empirical functions describing survival rates in the Preston data. S(t) is the cumulative survival rate at the end of each interval of time. It represents the product of the probabilities of survival up to and including the current interval. The function f(t) is the estimated probability per unit time of arrest for a felony ocurring within the interval (i.e., the unconditional failure rate). This value is computed by dividing the number of failures in the interval beginning at time t by the product of the total number of individuals (here, 1699) and the length of the interval (here, one year). Finally, h(t) is the hazard rate or conditional failure rate. It is computed as:

$$\frac{\text{number of failures in interval}}{(\text{interval length})(\text{number of survivors} - 1/2\ (\text{number of failures}))}$$

Table P

Empirically estimated survival function, probability density, and hazard rate

Full sample (N=1699)

Interval starting time (in years)	S(t)	f(t)	h(t)
0	.6518	.3479	.4215
1	.4695	.1807	.3252
2	.3518	.1148	.2868
3	.2897	.0594	.1935
4	.2490	.0383	.1510
5	.2242	.0230	.1051
6	.2105	.0124	.0627
7	.1927	.0159	.0887
8	.1865	.0053	.0323
9	.1802	.0053	.0345
10	.1734	.0053	.0382
11	.1705	.0018	.0172
12	.1705	.0000	.0000
13	.1667	.0006	.0225
14	.1667	.0000	.0000

All three functions indicate that indeed failure does occur primarily early on in follow-up period, with the probability of failing in each interval, given survival to the beginning of that interval, decreasing virtually monotonically. Appendix F presents similar life tables for the other operationalizations of failure.

B. Logit Analysis.

1. Model Construction.

As stated previously, the logit model is a static model. Consequently, the criterion variable (failure 3) must occur some time within a fixed window period. This window period in many studies is defined as being one to three years. One reason for this is that these studies are interested in relatively short term effects of some "treatment". The other reason, of course, is the expense involved in collecting data covering a long period of time for study populations of substantial size. The present analysis takes advantage of the excellent work of CYA's Rudy Haapanen in assembling a data base that followed up Preston parolees for more than a decade. We chose an eight year window

period for this study as a compromise in that it provides a fairly long period of time at risk without at the same time excluding too many subjects from the analysis due to early censoring. However, by doing this we make the assumption that, given two individuals A and B, if A is first arrested for a felony two months after his parole and B is first arrested seven and a half years after his parole, A is equivalent to B. Cases failing after the window period are necessarily regarded as successes (see Maltz, 1984 for a discussion).

Three mutually exclusive, random samples were drawn from the total analysis sample of 1633. Table Q gives the means and standard deviations for the three samples.

Table Q

Means and standard deviations for the three reduced samples

	A	B	C
N =	546	554	533
Failures =	453	456	437
Variable	x (sd)	x (sd)	x (sd)
failure 3	0.830	0.823	0.820
race	0.491	0.507	0.465
commitment type	3.106 (1.095)	3.083 (1.086)	3.081 (1.090)
priors arrests	1.390 (1.742)	1.341 (1.768)	1.435 (1.754)
age at parole	17.595 (1.126)	17.572 (1.125)	17.523 (1.059)
age first arrest	15.652 (1.727)	15.606 (1.768)	15.602 (1.594)

Initially, logit models making use of all five variables were estimated separately on the three samples. Table R presents the results.

Table R

Initial Logit Estimates

Sample	A	B	C
N	546	554	533
Failers	453	456	437
	Coefficient Values (t statistics)		
Variable			
constant	5.352 (2.63)	4.464 (2.26)	9.749 (4.24)
race	-0.607 (2.40)	-0.743 (3.00)	-0.285 (1.17)
commitment type	-0.056 (0.48)	0.012 (0.11)	0.085 (0.79)
prior arrests	0.322 (2.80)	0.364 (3.14)	0.496 (4.07)
age at parole	-0.019 (0.15)	-0.048 (0.36)	-0.472 (3.26)
age first arrest	-0.204 (2.15)	-0.130 (1.40)	-0.036 (0.37)
LL	-228.87	-238.59	-229.14
x^2	40.66	39.88	44.41
df	5	5	5
p	1.11E-7	2.07E-8	1.60E-7

As explained previously, samples A and B were then used as joint construction samples to identify mutually consistent variables with parameters of the final model estimated on the combined A and B samples. The decision rules built into the analytic scheme eliminated commitment type and age at parole from the model. The results are shown in Table S.

Table S

Logit estimates: Final Models

Sample	A	B	A+B
N	546	554	1100
Failers	453	456	909
	Coefficient Values (t statistics)		
Variable			
constant	4.968 (3.72)	3.991 (3.20)	4.437 (4.87)
race	-0.634 (2.57)	-0.743 (3.07)	-0.694 (4.03)
prior arrests	0.314 (3.05)	0.347 (3.30)	0.332 (4.55)
age first arrest	-0.210 (2.59)	-0.150 (1.97)	-0.178 (3.18)
LL	-228.99	-238.66	-467.86
x^2	40.42	39.74	79.8
df	3	3	3
p	1.01E-8	1.34E-8	1.55E-9

Generally, one can conclude from the model of Table S that the odds of being arrested for a felony within eight years subsequent to parole are greater for nonwhites, those with more priors and those who had their first first official contact at an early age. The individual effects of the variables, though, may prove more informative. Recall that this can be determined by examining the ratio of odds for two individuals who differ only in their value on the mth variable. Thus, given two individuals who differ only in regard to race, the odds of failing within eight years for nonwhites is approximately two times the odds for whites. In the case where two individuals differ only in regard to the number of priors with one having one prior arrest while the other has five, the odds of failing for the youth with five priors is approximately four times the odds of the individual with one prior arrest. (The odds ratio is 1.79 to 1 if the difference is taken as 1.8 -- the standard deviation of the variable in the data.) If two individuals differ only in regard to the age of onset as measured by the first official arrest with individual i beginning at age nine and person j beginning at age fifteen, the odds of failing within the window period for the individual who began at age nine are approximately three times those of the individual whose first arrest was six years later. (The odds ratio is 1.37 to 1 if the difference is 1.7 years, the standard

deviation of the age at first arrest.) Finally, if i is nonwhite, has six priors and was first arrested at age nine and j is a white with one prior and was arrested for the first time at age fifteen, person i's odds of being arrested for a felony within eight years after parole are approximately twenty two times j's odds.

2. Model Validation.

The solution model (A + B) was then applied to the data of sample C. The range of model-assigned probability values was .592 to .997 and this was divided into five non-overlapping and equal length segments. Table T presents the observed and expected numbers of failures in sample C when coefficients from the construction model are imposed.

Table T

Validation on sample C of logit model estimated on sample A+B

upper p in segment	segment n	failures observed	railures expected	std dev
0.673	19	12	11.95	2.10
0.754	103	73	72.56	4.62
0.835	143	111	115.14	4.73
0.916	134	116	117.60	3.79
0.997	134	125	126.88	2.58

Chi-square equals 1.478 with 9 degrees of freedom. The associated probability under the null hypothesis is .997. Further, the difference between observed and expected failures never exceeds one standard deviation in any segment.

C. Hazard Models.

1. The Proportional Hazard Model.

a. Model Construction.

Unlike the logit model, hazard models consider the time to failure or censoring to be an essential component of the model. Recall that the model being investigated in this paper takes the proportional hazard form:

$$\ln h(t,\underline{z}) = (\underline{z}'\underline{c} + b_1 t + a_1 \ln t)$$

Here $\underline{c}$ is a vector of coefficients; a_1 and b_1 are constants. All cases (N=1699) can be used at least in model construction. Table U presents the means and standard deviations for the three samples.

Table U
Means and standard deviations for three sub-samples based on the full sample (N=1699)

Sample	A	B	C
N =	571	574	554
Failures =	464	466	447
Variable	x (sd)	x (sd)	x (sd)
failure 3	0.813 (0.391)	0.812 (0.391)	0.807 (0.395)
race	0.487	0.503	0.473
commitment type	3.084 (1.107)	3.098 (1.080)	3.092 (1.089)
prior arrests	1.356 (1.730)	1.301 (1.752)	1.386 (1.740)
age at parole	17.597 (1.134)	17.564 (1.114)	17.534 (1.062)
age first arrest	15.658 (1.751)	15.634 (1.763)	15.628 (1.611)

Following the scheme used with the logit analysis, a solution model was obtained, the results of which are presented in Table V. As can be seen, the proportional hazard retains the same variables as did the logit model with estimated effects showing a qualitatively similiar relationship to failure.

Table V

Proportional Hazard Solution Model

Sample	A	B	A+B
N	571	574	1145
Failers	464	466	930
	Coefficient Values (t statistics)		
Variable			
Time independent term: c			
Constant	0.867 (1.92)	0.388 (0.92)	0.608 (1.99)
race	-0.338 (3.48)	-0.440 (4.63)	-0.392 (5.76)
prior arrest	0.117 (4.68)	0.161 (6.44)	0.138 (7.67)
age first arrest	-0.091 (3.25)	-0.058 (2.23)	-0.073 (3.84)
time coefficient b_1	-0.326 (8.63)	-0.352 (8.88)	-0.339 (12.84)
Ln time coefficient a_1	0.168 (2.71)	0.222 (3.40)	0.193 (4.33)
LL	-992.05	-997.82	-1991.05
x^2	299.83	315.55	613.10
df	5	5	5
p	1.55E-9	1.55E-9	1.55E-9

In sample A the hazard function passes through its maximum at six months, while sample B's maximum hazard is at about seven and a half months. For the combined sample, the hazard function maximum is about seven months after parole. That is, the risk of recidivism is rising for the first seven months and then gradually declines.

Generally, one may conclude from the solution model that, given survival to time t, nonwhites, those with more priors and those younger at first arrest are more likely to fail in the interval t to t+dt. Using the example where we have two individuals who

differ only on the value of one variable (i.e., white vs. nonwhite, one prior vs. five priors, nine at first arrest vs. fifteen at first arrest), we are able to assess the relative strength of the variables. For the proportional hazard model these ratios of risks of near term failure remain constant in time. With regard to race, the risk of recidivism per unit time for a nonwhite is about one and a half times that of a white. An individual who differs from another only in that he has five prior arrests while the other just has one is 1.74 times more likely to fail. And an individual who was first arrested at age nine poses about one and a half times the risk of failure of the individual who was fifteen. Consequently, a nonwhite with five priors who was first arrested at age nine is about four times more likely to be arrested for a felony in each time interval subsequent to parole (given survival to the beginning of the interval) than the white with one prior and age fifteen when first arrested.

b. Model Validation.

Table W asseses the final model as a "predictor" using the data of sample C.

Table W

Validation on sample C of Proportional Hazard solution model built on sample A + B

upper p in segment	segment n	failures observed	failures estimated	std dev
0.687	69	44	45.36	3.94
0.765	100	78	72.31	4.46*
0.844	127	102	102.81	4.42
0.922	140	122	123.13	3.84
1.000	97	91	92.92	1.97

* = Estimated and observed failures differ by more than one standard deviation in this segment.

This division of the probability range into five segments produces a chi-square of 2.80 with 9 degrees of freedom. Under the hypothesis that the pattern of observed and expected outcomes are from the same population, the associated probability is .972.

2. The "Full" Hazard Model.

a. Model Construction.

As stated above the proportional hazard restricts the covariate coefficients of the time dependent terms to zero. Alternatively, it may be plausible that covariate effects change over time as defined by the model

$$\ln h(\underline{z},t) = \underline{z}'(\underline{c} + \underline{a} \ln (t) + \underline{b}t)$$

Table X presents the the solution obtained in the usual way by using samples A and B as the data for model construction.

Table X

Full Hazard Solution Model

Sample	A	B	A+B
N	571	574	1145
Failers	464	466	930
	Coefficient Values (t Statistics)		
Variable			
Time independent term: $\underline{c}$			
race	-0.335 (3.45)	-0.425 (4.38)	-0.381 (5.60)
priors	0.114 (3.93)	0.147 (5.25)	0.131 (6.55)
parole age	0.047 (1.57)	0.051 (1.82)	0.047 (2.35)
age first arrest	-0.089 (2.78)	-0.089 (2.87)	-0.087 (3.95)
Linear term in time: $\underline{b}$			
commitment type	-0.023 (1.21)	-0.024 (1.26)	-0.023 (1.64)
parole age	-0.015 (3.75)	-0.017 (4.25)	-0.016 (5.33)
Ln term in time: $\underline{a}$			
race	0.029 (0.37)	0.058 (0.72)	0.043 (0.77)

(Table X continued)

commitment type	0.040 (1.00)	0.047 (1.09)	0.043 (1.48)
parole age	-0.013 (0.68)	-0.024 (1.20)	-0.018 (1.29)
age first arrest	0.018 (0.86)	0.032 (1.45)	0.024 (1.60)
LL	-991.32	-994.92	-1987.26
x^2	301.29	321.36	620.69
df	9	9	9
p	1.55E-9	1.55E-9	1.55E-9

Unlike the logit and proportional hazard models, all variables are retained. In the final model, race, commitment type, parole age and age at first arrest appear to be changing in effect over time, while prior arrest effects remain constant. Retaining the example used in both the proportional and logit models, the relative effects of the variables may be illustrated as follows:

Race:

$$\ln h(i)/h(j) = -1*(-.381 + .043*\ln(t))$$

This function decreases monotonically in time but very slowly after the first month or so. Again it should be remembered that it estimates a near term probability of failure, given that the subject has survived to time t. With two individuals who are indentical except for their race, by the end of the first year the nonwhite is approximately one and a half times more likely to be arrested for a felony in the near future than is the white. By year ten the hazard ratio has decreased to 1.3. Recall that with the proportional hazard, nonwhites in every interval were approximately one and one half times more likely to fail. A likely interpretation of this weak time dependence is that the model has picked up from the data a signal of a relatively stronger risk differential based on race in the first few months after release, followed by a risk ratio that remains relatively constant in time. Indeed, the data shows the overall ratio of non-white to white failure rates to be 1.12 when calculated for the entire follow-up period. For the first six months at risk, however, this ratio is 1.51.

t (years)	h(i)/h(j)
1	1.46
2	1.42
5	1.37
8	1.34
10	1.33

Commitment type:

$$\ln h(i)/h(j) = -3*(-.023*t + .043*\ln(t))$$

The function in parentheses has the typical gamma density shape, starting at zero, passing through a maximum and then returning asymptotically to zero. With the -3, representing the difference in commitment type between two subjects, of course, the function maps into a U-shaped curve in time. Thus, if person i was committed for a violent offense and person j for a minor offense, the individual with the more serious committing offense is approximately 1.07 times more likely to be arrested for a felony at the end first year. The function passes through its minimum at about 22 and a half months. By year ten, the individual with a violent committing offense is approximately one and a half times more likely to be arrested for a felony given survival to the beginning of year ten.

t (years)	h(i)/h(j)
1	1.07
1.86	1.049
2	1.05
5	1.15
8	1.33
10	1.48

One might surmise that the U-shape is the model's attempt to represent a hazard ratio that increases slowly but monotonically in time, starting from a finite value.

Prior arrests:

$$\ln h(i)/h(j) = 4*(.131)$$

Given two individuals who differ only in the number of prior official arrests, where individual i has one prior and person j has five, the individual with four more priors is 1.7 times more

likely to fail in any interval given survival to the beginning of that interval relative to the individual with just one prior. This result differs little from that estimated with the proportional hazard model.

Parole age:

$$\ln h(i)/h(j) = 2*(.047 - .015*t - .018*\ln(t))$$

Having two individuals differing only in their age at the time of parole from Preston, where person i was 17 and person j was 15, results in a decreasing function in time. After one year the 17 year old at parole is 1.07 times more likely to fail but this ratio decreases over the ten year period. For instance, by the end of year two they have approximately equal probabilities of failing given survival. However, after eight years at risk, the individual who was 15 at parole (j) is about 1.3 times more likely to fail relative to the person who was 17 (i). The variation in the ratio is shown in the table below:

t (years)	h(i)/h(j)
1	1.06
2	1.01
3	0.96
4	0.92
5	0.88
6	0.85
7	0.82
8	0.79
9	0.76
10	0.74

Age at first arrest:

$$\ln h(i)/h(j) = -6*(-.087 + .024*\ln(t))$$

Where person i was 9 and individual j was 15 at their first officially recorded arrest, at t = 1 year after release the youth who was 9 at first arrest is about 1.7 times more likely to fail. The ratio decreases monotonically over time. By year 10 it has the value 1.21. Recall that the proportional hazard model estimated that the individual who was 9 at first arrest was constantly about one and a half times more likely to be arrested for a felony than the youth who was 15.

t (years)	h(i)/h(j)
1	1.69
2	1.52
5	1.33
8	1.25
10	1.21

b. Model Validation.

Validation on sample C of the model built on the combined sample A + B gives the results shown in Table Y.

Table Y

Validation on Sample C of full hazard solution model estimated on sample A + B

upper p in segment	segment n	failures observed	failures expected	std dev
0.682	55	36	36.23	3.51
0.761	108	82	77.81	4.66
0.841	134	105	108.10	4.56
0.920	141	126	123.75	3.88
1.000	95	88	90.70	2.02*

*= Estimated and observed failures differ by more than one standard deviation in this segment.

With 9 degrees of freedom the chi-square is 3.38 with a probability under the null hypothesis of .947.

Finally, we may assess the adequacy of the model built on the combined data set A+B for "predicting" failure within time intervals. Table Z compares the observed and predicted failures for successive 13 week intervals spanning eight years of risk. In addition, the predicted average hazard rate for both successes and failures are given for each interval. As in the similar analysis of the North Carolina data, we have here combined the observations from all three samples.

Table Z
Predicted and expected failure among population
at risk at 13 week intervals
(N = 1699)

t_1 (days)	$N(t_1)$	Obs fail	Exp fail (sd)	Av. $P(f_{ail})$ (sd) x100	Av. $P(s_{urv})$ (sd) x100	p(t)
1	1699	149	155.4 (11.7)	12.14 (6.37)	8.86 (4.70)	.000
92	1548	194	160.7** (11.9)	12.32 (6.05)	10.10 (4.41)	.000
183	1354	138	139.31 (11.07)	11.42 (4.74)	10.16 (4.12)	.001
274	1215	110	121.5* (10.37)	11.26 (4.89)	9.88 (3.76)	.002
365	1105	92	105.7* (9.70)	10.67 (3.90)	9.47 (3.49)	.002
456	1008	92	91.7 (9.07)	10.26 (3.92)	8.98 (3.19)	.001
547	911	68	77.9* (8.39)	9.98 (3.95)	8.44 (2.88)	.001
638	841	54	67.3* (7.83)	9.34 (3.43)	7.91 (2.61)	.001
729	784	65	58.60 (7.33)	8.69 (2.77)	7.36 (2.37)	.000
820	718	51	49.8 (6.78)	7.78 (2.09)	6.87 (2.21)	.001
911	665	46	42.9 (6.31)	7.04 (2.53)	6.41 (2.02)	.05
1002	615	33	37.0 (5.88)	6.64 (1.75)	5.98 (1.88)	.02
1093	578	32	32.3 (5.51)	5.75 (1.74)	5.58 (1.75)	.29
1184	543	25	28.3 (5.17)	5.81 (1.41)	5.18 (1.64)	.02

(Table Z continued)

1275	516	21	25.0 (4.86)	4.78 (1.70)	4.85 (1.51)	.57
1366	492	23	22.2 (4.59)	4.66 (1.55)	4.51 (1.40)	.32
1457	468	18	19.7 (4.33)	4.54 (0.96)	4.19 (1.31)	.07
1548	448	18	17.4 (4.09)	3.89 (1.07)	3.90 (1.22)	.51
1639	425	19	15.47 (3.84)	4.05 (1.48)	3.60 (1.11)	.10
1730	406	10	13.6 (3.61)	3.25 (0.84)	3.35 (1.04)	.63
1821	395	11	12.3 (3.44)	3.57 (1.01)	3.09 (0.96)	.06
1912	384	9	11.0 (3.27)	3.17 (1.13)	2.86 (0.88)	.20
2003	372	11	9.9 (3.10)	3.05 (0.95)	2.64 (0.81)	.08
2094	360	9	8.8 (2.93)	2.77 (0.64)	2.44 (0.76)	.07
2185	348	8	7.9 (2.77)	2.55 (0.61)	2.25 (0.70)	.08
2276	339	5	7.1 (2.63)	2.34 (0.51)	2.08 (0.66)	.13
2367	332	6	6.4 (2.50)	1.99 (0.63)	1.92 (0.61)	.40
2458	325	2	5.8* (2.38)	1.64 (0.34)	1.78 (0.57)	.71
2549	323	5	5.3 (2.29)	1.52 (0.48)	1.65 (0.53)	.73
2640	316	9	4.8* (2.18)	1.82 (0.84)	1.52 (0.45)	.14
2731	306	6	4.3 (2.06)	1.41 (0.20)	1.41 (0.45)	.50

(Table Z continued)

2822	297	7	3.9* (1.95)	1.23 (0.21)	1.30 (0.42)	.81

* = Expected and Observed failures differ by more than one standard deviation.
** = Expected and Observed failures differ by more than two standard deviations.

A comparison of the sequences of observed and expected failures seems to indicate that from years 3 to 7 the model is doing a pretty good job of tracking the changing probabilities over time. Initially, its performance is more erratic -- especially in the second quarter following parole.

The sequences of hazard scores for survivors and failures are, perhaps, more interesting. For the survivors, the average scores increase for the first three quarters of a year and then decrease monotonically. For the failures the average score peaks at the second quarter. Then with the exception of the periods beginning at times 1184, 1639, 1821 and 2640 days, it decreases steadily throughout the remainder of the eight years.

For the first three years the differences between failures' and survivors' average scores are statistically significant. After the interval beginning on day 1093 the significance test results become quite variable; and, indeed, in 6 of the last 20 intervals the failures' averages are less than the survivors'.

Finally, the decreasing standard deviations of both sequences of scores should be noted.

Qualitatively, at least, this seems to be saying that the model is doing pretty much what we expect of it. Conditional probabilities of failure are changing over time and in each interval a greater fraction of the high scorers are being dropped from the population as failures. Gradually, the very highest risk individuals have either been weeded out or they have survived long enough to be well beyond their time of greatest risk. The population of scores thus tends to become more and more homogeneous.

Appendix A

Offenses used for construction of arrest for a felony and arrest for a felony resulting in conviction.

1) murder1
2) murder2
3) manslaughter
4) assault-felony
5) forcible rape
6) other crimes against person
7) bank robbery
8) armed robbery
9) strong armed robbery
10) burglary
11) forgery
12) grand theft
13) arson
14) buying and receiving
15) auto burglary
16) other felony theft
17) grand theft auto

Appendix B

Fraction failing; means and standard deviations of observation time for other failure operationalizations.

N = 1699
Time = days to failure or censoring

Variable	x	sd	min	max
failure1	0.929			
time1	575.806	1044.034	6	5156
failure2	0.902			
time2	769.034	1217.570	6	5191
failure4	0.722			
time4	1743.180	1665.606	11	5258

Appendix C

1) murder 1
2) murder 2
3) manslaughter
4) assault-felony
5) forcible rape
7) assault-misd
8) person-other
10) robbery-bank
11) robbery-armed
12) robbery-strongarm
13) burglary
14) trespass
15) receiving stolen property
16) forgery
17) theft-grand
18) theft-petty
19) shoplifting
20) arson
21) malicious mischief
22) burglary-auto
25) child molestation
26) other felony theft
27) other misd theft
29) statutory rape
30) homosexual relations
31) incest
32) prostitution-solicitation
33) other sex crimes
34) grand theft auto
35) joyriding
36) hit and run
37) traffic
38) other auto violations
39) weapon,possession
40) resisting,obstruction
41) loitering,vagrancy
42) disturbing the peac
43) gambling
44) parole violation
45) probation violation
46) game violation
47) other local codes
48) public safety
49) suspicion of felony
50) suspicion of misd
51) aiding and abetting
52) other non-status
53) drunk
54) drunk driving
56) other liquor
57) drug sale-heroin,coke,morphine
58) drug sale-LSD,hallucinogenics
59) drug sale-pot,hashish
60) drug sale-pills
61) other sale,manufacturing
62) drug use-herion,coke,morphine
63) drug use-LSD,hallucinogenics
64) drug use-pot,hashish
65) drug use-pills
66) drug use-sniffing
67) other use or possession
68) drugs and driving
69) drug,situational violations
70) suspicion of drug use
71) other misc drug
73) runaway
76) missing person
78) truancy
80) curfew
81) beyond control
82) possession of alcohol
84) violation of juvenile probation
85) failure to appear
86) escape
89) other status offenses
90) held for other jurisdiction
91) family dispute
92) no precipitating offense
93) missing child
94) no offense given
95) neglect, abused
96) expelled from home
97) attempted suicide
98) nonspecified offense

Appendix D

Conviction was considered to have occurred if any of the following resulted.

1) dismissed, convicted of other charge
2) suspended sentence
3) convicted, sentence unknown
4) fine or restitution
5) work project
6) probation without wardship
7) probation with wardship
8) adult probation
9) county juvenile
10) jail
11) California Rehabilitation Center
12) California Youth Authority
13) California Department of Corrections
14) Non-California prison
15) death penalty

Appendix E

Means and standard deviations for successes and failures

	failure1 mean(sd)	failure2 mean(sd)	failure4 mean(sd)
race			
failures	0.481	0.478	0.452
successes	0.583*	0.575*	0.581*
commitment type			
failures	3.10 (1.09)	3.09 (1.09)	3.09 (1.08)
successes	3.03 (1.14)	3.13 (1.11)	3.10 (1.12)
age at parole			
failures	17.55 (1.11)	17.54 (1.10)	17.52 (1.08)
successes	17.84 (1.06)*	17.80 (1.11)*	17.67 (1.16)*
prior arrests			
failures	1.42 (1.77)	1.44 (1.77)	1.55 (1.80)
successes	0.36 (0.84)*	0.47 (1.06)*	0.82 (1.45)*
age first arrest			
failures	15.57 (1.69)	15.55 (1.69)	15.45 (1.68)
successes	16.55 (1.74)*	16.47 (1.70)*	16.13 (1.71)*
time (in days)			
failures	350 (499)	438 (542)	894 (852)
range	6 - 4749	6 - 4207	11 - 4726
successes	3541 (1647)	3804 (1477)	3944 (1170)
range	44 - 5156	44 - 5191	44 - 5258
Number of cases	1699	1699	1699
failures	1579	1532	1226
successes	120	167	473

* significantly different: p less than .05

Appendix F

Estimated survival, probability density and hazard rates

interval starting time (years)	failure1			failure2			failure4		
	S(t)	f(t)	h(t)	S(t)	f(t)	h(t)	S(t)	f(t)	h(t)
0	.3219	.6775	1.026	.4274	.5721	.8023	.7443	.2554	.2931
1	.1668	.1530	.6349	.2406	.1848	.5592	.6061	.1371	.2047
2	.1196	.0453	.3298	.1746	.0642	.3182	.4901	.1136	.2116
3	.0969	.0212	.2093	.1382	.0347	.2327	.4183	.0695	.1582
4	.0873	.0088	.1049	.1225	.0147	.1202	.3670	.0489	.1305
5	.0820	.0047	.0623	.1117	.0100	.0924	.3345	.0306	.0928
6	.0732	.0077	.1135	.1032	.0077	.0785	.3149	.0182	.0604
7	.0676	.0047	.0796	.0973	.0053	.0596	.2924	.0206	.0739
8	.0669	.0006	.0108	.0945	.0024	.0285	.2800	.0112	.0433
9	.0638	.0024	.0462	.0896	.0041	.0536	.2706	.0082	.0341
10	.0615	.0018	.0380	.0881	.0012	.0167	.2641	.0053	.0244
11	.0597	.0012	.0290	.0872	.0006	.0101	.2611	.0018	.0116
12	.0597	.0000	.0000	.0872	.0000	.0000	.2573	.0012	.0147
13	.0566	.0006	.0526	.0872	.0000	.0000	.2573	.0000	.0000
14	.0566	.0000	.0000	.0872	.0000	.0000	.2573	.0000	.0000

VI. SOME CONCLUDING REMARKS.

The work reported here was begun as an exploration of where certain analytic methods might lead in the study of recidivism considered as an inherently stochastic process. This final section of what is already an overly long paper will attempt to summarize very briefly what we feel are some of the more interesting things we've noted along the way.

1. On the Prediction of Individual Failures.

We think that the analyses clearly support the concept of "recidivism" as a process in which chance plays an essential role. Of course, this is no proof. Some different set of explanatory variables or different mathematical forms could conceivably produce results that reliably separate a study population into distinct groups: almost certain successes and almost certain failures. But the variables and relatively simple mathematical forms used here assign to most subjects failure probabilities that lie in some middle ground -- between .10 and .90 in the two year follow-up of North Carolina subjects; between .60 and .90 in the 8 years of the CYA data.

The implication, of course, is that analyses that force a population division into predicted successes and failures must contend with substantial and relatively irreducible rates of prediction error -- a result that will hardly astonish anyone with even the most cursory familiarity with the prediction literature. If prediction instruments are to be useful in practice as guides for dichotomous decisions on individual dispositions, it is essential that they also take into account the benefits and costs associated with their expected rates of right and wrong decisions. With failure probabilities distributed continuously over a wide range, these rates can be quite sensitive to the value chosen as the boundary between the "good" and "bad" risks. That choice seems to us to be an important matter for policy makers.

2. On a More Direct Policy Use of Failure Probabilities.

Criminal justice policy makers are being compelled nowadays to experiment with a variety of non-incarcerative sanctions, with an attendant concern over the risks to public safety entailed in leaving convicted offenders more or less free in society. It is in the "more or less" aspect of these sanctioning policies that actuarial models would seem to find a natural role. To the extent that non-incarcerative sanctions can still impose some variability in restricting offenders' freedom, these models offer a basis for the differential allocation of criminal justice resources available for supervision and surveillance.

We would argue as well that it is here that hazard models offer a practical advantage over static models such as the logit. If a term of probation or parole supervision is of any substantial duration, it can be assumed that an agency would, as time goes on, "reclassify" individuals who are still under their official charge. Changes in risk assessment over time are precisely what hazard models would purport to be able to do.

3. On the Importance of Demonstrations of Model Validity.

In a sense the general analytic plan used in this study considered sub-samples from a population as the units of analysis. This was done in an attempt to build into the investigation some rudimentary requirement of model replicability. We think this deserves further research consideration. Given a data base of sufficient size, the question is whether we can get more useful information by studying it as a collection of independent sub-samples or by analyzing all the data at once.

The authors don't have an answer to which procedure is better. But we would point out that the question is not simply a technical one. In criminal justice policy applications it can be assumed that there is some rough order of population size for which a model must give reliable results -- something, perhaps, like the number of "predictions" that would have to be made over a period of six months or a year. It seems essential to give potential users of a model some practical sense of its accuracy when applied to samples of that size. Statistical assurances that the model will prove to be valid over the long run may not be convincing and, indeed, could be misleading. We recommend that, whenever possible, some empirical demonstration of validity be built into the design of all studies that would draw conclusions and make policy recommendations based on statistical models.

4. Reporting Probability Distributions.

Virtually all attempts to synthesize some field of criminal justice research and infer what cumulative progress has been made lament the lack of comparability among studies investigating similar questions. In this regard researchers using an actuarial model to study recidivism could, we think, take a small and relatively easy step in the right direction by reporting how their model distributes the probability of failure among the study population -- not just the model results. In the same vein studies evaluating a treatment program or policy innovation should report the change found in the probability distribution rather than just their determination of whether or not a significant effect was found.

5. The North Carolina Results.

In the North Carolina data three variables were retained that the authors find particularly interesting.

First, the rate of in-prison rule violations turned out to have a surprisingly strong relation to failure. The literature on this gives rather more mixed results. We would certainly encourage further investigation of this variable in future recidivism modeling studies.

Second, the crime clearance rate in a subject's jurisdiction of release is quite obviously related to his recidivism risk when "failure" is observed only through some criminal justice system action. This need not have anything at all to do with theories of deterrence. If this variable is not included in a model, the analyst is making the assumption that, given a crime, the probability of arrest (or conviction, or return to prison) is uniform across the jurisdictions represented in his study population. The results reported here suggest that it might well be worthwhile to test that assumption.

Third, the fact that a jurisdiction's unemployment rate seems to have a fairly stable if not particularly powerful relation to recidivism in the models developed here warrants further research attention. As noted earlier, the unemployment rate seems to us to be standing as a surrogate for a more complex set of social conditions. For it is difficult to conceive of a process in which a higher unemployment rate would directly produce a lower recidivism probability. Recidivism modeling studies might, therefore, be recommended in which some more extended set of measures of local socio-economic conditions are tested.

Something might also be said about some of the variables that were not retained -- the modeling algorithm's way of saying it could find no consistent relation with recidivism probability. In particular, the three variables indicating whether the subject had been on work release, whether he had participated in educational or vocational training programs and whether he had been involved with prison industries or been assigned prison duties were all dropped from both the logit and hazard models.

Theoretically, these variables were intended as measures of some improvement in the individual's competence in the job market after release from prison. The disappointing result here is not inconsistent with Orsagh and Marsden's conclusion based on a very different set of analyses. They find only weak evidence for a main effect of these variables but rather more encouraging results in analyses of interactive effects.

6. The Preston Results.

What is perhaps most surprising is that models built on such a small number of covariates and spanning such a long period of time would fit the validation sample outcomes at all. Be that as it may, comparisons between the North Carolina and the Preston results would seem to give some clues for general improvement in such models.

The covariate-rich North Carolina data base results in models that distribute individual failure probabilities over a very wide range. The parsimonious data set used for the Preston analyses result in models that see the population as much more homogeneous. And unlike the North Carolina models, the proportional hazard model here does about as well as the fully parametrized form. Is the recidivism process really fairly uniform and simple among the Preston youths and quite different from the process operating among adult offenders in North Carolina?

From Table A it might be noted that the two year failure rate among North Carolina releasees was about .48. Table P gives the two year rate for the Preston sample as .53. The somewhat higher value might well be "explained" in terms of differences in average ages of the two populations.

Even more striking, however, is the comparison of Table Z with the table given in the Appendix to the North Carolina section of the paper. Among those who <u>failed</u> in each 13 week interval, the vectors of average failure probabilities assigned by the hazard models are quite close during the first two years at risk. The average failure probability of <u>survivors</u>, however, is systematically smaller in the North Carolina than in the Preston results -- at least up to the eighth quarter year. This suggests to us that the Preston models are overly parsimonious -- that the use of a more extensive set of individual characteristics might have resulted in a reliable model that, at least over the near term following release, discriminates better between subjects' recidivism risks. Indeed, it might be the case that short and long term failure probabilities depend on somewhat different sets of subject characteristics.

7. Further Research.

The logit model was found to perform somewhat better in assigning individual probabilities of failure within some fixed period of time. These results suggest that it might be preferred as a basis for informing dichotomous decisions since the characteristic decision criterion in that case typically is concerned with relative risk over some policy-determined period

following release.

But hazard models quite generally possess practical as well as analytic advantages. Their case is eloquently argued by Maltz (1984). In evaluation studies hazard models might uncover an effect of delayed time to failure that would be meaningful to a policy decision but would be completely masked in the results obtained from a fixed-time model. And, as illustrated in this paper, the time dependent assignment of risk scores could find application in the allocation of parole or probation supervision resources.

We ourselves found particularly interesting the latitude the hazard model results give for speculation about why the recidivism probability evolves in time the way a model says it does. Certainly, caution and a very generous amount of skepticism are to be advocated in interpreting modeling results in such detail -- especially when considering the isolated contributions of particular variables as reflecting potential causes of failure.

But it is not entirely fanciful, we think, to look on hazard models as a sort of dynamic counterpart of fixed-time models. It would not require much of an extension of the hazard modeling formalism used here to allow the explanatory variables to change in the course of time -- basing risk assessments on current values of individual and contextual variables rather than simply on the information available at the time of release. A further extension of the formalism could also allow for multiple failures, offering a method for examining the sequence of events that go to make up a criminal career.

Given the ready accessibility and the power of today's computers, recidivism research is no longer much constrained by the costs of parameter estimation, even for models that would investigate questions of considerable analytic complexity. Of course, the large data bases needed for building and testing such models are expensive and time consuming to assemble. Slowly but steadily, however, such data bases are becoming more readily available for modeling research.

References

Aldrich, J. H., and F. D. Nelson
1984 Linear Probability, Logit, and Probit Models. Beverly Hills, CA: Sage Publications

Barnett, A., A. Blumstein and D. P. Farrington
1987 "Probabilistic Models of Youthful Criminal Careers." Criminology 25: 83-107.

Box, G. E. P., and G. M. Jenkins
1976 Time Series Analysis. San Francisco, CA: Holden-Day.

Copas, J. B.
1985 "Prediction Equations, Statistical Analysis and Shrinkage." Pp. 232-257 in D. P. Farrington and R. Tarling, eds., Prediction in Criminology. Albany, NY: State University of New York Press.

Copas, J. B., and R. Tarling
1986 "Some Methodological Issues in Making Predictions." Pp. 291-313 in A. Blumstein, J. Cohen, J. A. Roth, and C. A. Visher, eds., Criminal Careers and "Career Criminals." Washington, DC: National Academy Press.

Duncan, O. D., L. E. Ohlin, A. J. Reiss, and H. R. Stanton
1952 "Formal Devices for Making Selection Decisions." American Journal of Sociology 58: 573-584.

Gottfredson, S. D., and D. M. Gottfredson
1986 "Accuracy of Prediction Models." Pp. 212-290 in A. Blumstein, J. Cohen, J. A. Roth, and C. A. Visher, eds., Criminal Careers and "Career Criminals." Washington, DC: National Academy Press.

Gottfredson, D. M., and M. Tonry, eds.
1987 Prediction and Classification: Criminal Justice Decision Making. Chicago, IL: The University of Chicago Press.

Haapanen, R. A., and C. F. Jesness
1982 Early Identification of the Chronic Offender. Sacramento: California Youth Authority.

Harris, C. M., A. R. Kaylan, and M. D. Maltz
1981 "Recent Advances in the Statistics of Recidivism Measurement." Pp. 61-80 in J. A. Fox., ed., Models in Quantitative Criminology. New York, NY: Academic Press.

Kalbfleisch, J. D., and R. L. Prentice
1980 The Statistical Analysis of Failure Time Data. New York, NY: John Wiley and Sons.

Larrimore, W. E.
1983 "Predictive Inference, Sufficiency, Entropy and an Asymptotic Likelihood Principle." Biometrika 70: 175-181.

Larrimore, W. E., and R. K. Mehra
1985 "The Problem of Overfitting Data." Byte October 1985: 167-180

Lee, E. T.
1980 Statistical Methods for Survival Data Analysis. Belmont, CA: Wadsworth.

Maltz, M. D.
1984 Recidivism. Orlando, FL: Academic Press.

Orsagh, T., and M. E. Marsden
1984 Rational Choice Theory and Offender Rehabilitation. Final Report to the National Institue of Justice. Washington, DC: National Criminal Justice Reference Service NCJ-95316.

Pritchard, D. A.
1979 "Stable Predictors of Recidivism: A Summary." Criminology 17: 15-21.

Reiss, A. J.
1951 "The Accuracy, Efficiency, and Validity of a Prediction Instrument." American Journal of Sociology 57: 115-120.

Schmidt, P., and A. D. Witte
1980 "Evaluating Correctional Programs: Models of Criminal Recidivism and an Illustration of Their Use." Evaluation Review 4: 585-600.

1984 An Economic Analysis of Crime and Justice. Orlando, FL:Academic Press.

1987 How Long Will They Survive?: Predicting Recidivism Using Survival Models. In Press.

Stollmack, S., and C. M. Harris
1974 "Failure Rate Analysis Applied to Recidivism Data." Operations Research 23: 1192-1205.

Waldo, G. P., and T. G. Chiricos
1977 "Work Release and Recidivism: An Empirical Evaluation of a Social Policy." Evaluation Quarterly 1: 87-108.