PREDICTING HIGH-RISK CHOLESTEROL LEVELS

BY

ALAN M. GARBER, RICHARD A. OLHSEN
and HEPING ZHANG

TECHNICAL REPORT NO. 151
JANUARY 1992

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANTS
5 R01 GM21215-17 AND 1 R01 CA55325

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Predicting High-Risk Cholesterol Levels

By

Alan M. Garber, Richard A. Olshen and Heping Zhang

Technical Report No. 151

January 1992

Prepared Under the Auspices

Of

Public Health Service Grants

5 R01 GM21215-17 and 1 R01 CA55325

Division of Biostatistics

Stanford University

Stanford, California
PREDICTING HIGH-RISK CHOLESTEROL LEVELS

by

Alan M. Garber
Department of Medicine
MSOB X-214
Stanford University School of Medicine
Stanford, California 94305-5479

and

Department of Veterans Affairs Medical Center
3801 Miranda Avenue
Palo Alto, California 94304

Richard A. Olshen
Division of Biostatistics
Stanford University School of Medicine
HRP Building, Room 110A
Stanford, California 94305-5092

Heping Zhang
Department of Epidemiology and Public Health

and

Department of Statistics
Yale University
New Haven, Connecticut 06510

November 24, 1991

This manuscript is based in part on Technical Report 140 of the Division of Biostatistics. Part of Section III, and all of Sections IV and V, are completely new.
SUMMARY

The pattern of longitudinal change in cholesterol levels has important implications for screening policies and for understanding the role of cholesterol as a risk factor for coronary heart disease. We explored a variety of longitudinal models to predict changes in cholesterol levels over a long period, emphasizing the probability that an individual will develop a cholesterol level that requires further diagnostic tests or treatment. After discovering that cholesterol levels are not first-order Markovian, we developed and applied several models to predict cholesterol levels, based on the pattern of previous cholesterol levels. We then used receiver-operating characteristic curves to summarize the sensitivity and specificity of prediction rules based on these models. The results of our analyses suggest that the models can be used to identify subgroups of individuals who are unlikely to develop very high-risk cholesterol levels within a two-year interval, and despite the failure of the Markovian hypothesis, good predictions can be based on a single recent measurement.

Key words: CHOLESTEROL; PREDICTION; BOOTSTRAP; CROSS-VALIDATION; MARKOV PROCESS; RECEIVER-OPERATING CHARACTERISTIC CURVES
I. INTRODUCTION

Several nations and professional organizations have initiated campaigns to prevent coronary heart disease by detecting and treating high blood cholesterol levels (European Atherosclerosis Society, 1987; National Cholesterol Education Program, 1988; U.S. Preventive Services Task Force, 1990). Many aspects of the cholesterol programs are controversial, such as the cholesterol level that requires treatment, the type of treatment that should be applied, and the groups that should be tested (Palumbo, 1989; Garber, 1989a; Garber, 1989b; Toronto Working Group on Cholesterol Policy, 1990). In contrast, there has been little disagreement, or even discussion, of how often the cholesterol level should be measured. We address this issue by examining the predictability of cholesterol levels and the information loss, if any, that would result from testing less frequently.

Related literature is either concerned with short-term fluctuations in measured cholesterol levels or with “tracking” – measuring the correlation between cholesterol levels drawn years apart. A number of factors contribute to short-term fluctuations. Laboratories that fail to standardize and calibrate their measurements adequately often measure cholesterol inaccurately (Laboratory Standardization Panel, 1988). Even if the laboratory is perfectly accurate, physiological phenomena can cause an individual’s cholesterol level to fluctuate. Stress, acute illness, seasonal effects, posture, and aspects of the blood-drawing technique can cause significant though transient changes in cholesterol levels (Thomas et al., 1961; Gordon et al., 1987; Hegsted and Nicolosi, 1987). Longer-term variation and issues of predictability are explored in a number of studies of tracking. Our effort is close in spirit to the latter literature. The tracking literature, however, primarily addresses whether a high cholesterol level at a young age is likely to predict the risk of cardiac disease in the distant future, while our work uses different methods and focuses on shorter-term prediction of cholesterol levels. In particular, our methods are designed to provide information about the performance of alternative policies toward cholesterol screening.

Although interest in testing strategies directs our efforts toward prediction, we also investigate the “natural history” of cholesterol levels. The data we use, from the 20-year followup of the Framingham Heart Study, are particularly suitable for this purpose because they cover a long time span, they include multiple measurements of cholesterol levels, and because most of the period of observation preceded the time when hypercholesterolemia
was deliberately and routinely treated with either diet or medications.

The consequences of changing the interval between screening tests depend in a direct manner on the predictability of the test results. The longer the time elapsed before a cholesterol test is repeated, the greater the chance that an abnormality will develop and the longer the time that it will remain undetected. But both inconvenience and costs diminish as the interval between tests is lengthened. If the cholesterol level varies little over time or changes in a predictable manner, the advantages of infrequent testing may outweigh the risk of prolonging the duration of untreated hypercholesterolemia.

There are other important reasons to understand the dynamics of the cholesterol level. For example, it may yield insights into the relationship between the variation in the measured level and the role of cholesterol as a risk factor. Some have argued (MacMahon et al., 1990; Collins et al., 1990) that random error in cholesterol measurements causes epidemiologic studies to understate the relation between the cholesterol level and the risk of coronary heart disease. If cholesterol levels vary widely and unpredictably, and if cardiac risk is related to a long-term average cholesterol level rather than the most recently measured value, a single cholesterol measurement will not predict risk accurately even if there is no measurement error. Similarly, prediction of future measurement values that is affine (linear with an intercept) in previous cholesterol measurements may be imprecise if cholesterol levels vary with time in a predictable but nonaffine fashion. Thus a second goal of our analysis is characterization of the predictable and unpredictable variability in cholesterol levels.

We assess the probability of a change from a "desirable" cholesterol level to an "elevated level" – i.e., a level considered high enough to warrant further testing or treatment – for various testing schedules. We address the clinical question that motivates our analysis: can we reliably distinguish those individuals who will develop hypercholesterolemia from those who will not? We answer this question by developing predictive models with regression techniques, including spline regression that is constrained in a nonstandard way. Markov and low order ARMA models are tested and rejected. We use receiver operating characteristic (ROC) curves to compare the performance of alternative prediction methods, and we develop Gaussian and nonparametric methods to compute confidence regions for points on the curves.
II. THE FRAMINGHAM DATA

The celebrated Framingham Heart Study has been one of the primary sources of information about the role of several risk factors in the development of heart disease. More than 5,000 men and women who were residents of Framingham, Massachusetts in the late 1940s and who were between 30 and 59 years old at that time entered the study (Dawber, 1980). Participants were examined every two years. Among the characteristics recorded at each examination were blood pressure, blood cholesterol, many other blood chemistry values, weight, and self-reported cigarette smoking. Several clinical outcomes were monitored, including symptoms of coronary heart disease, myocardial infarction, death from coronary heart disease, and death from all causes. Because only 3,172 of the participants had a cholesterol measurement on the first examination, for our analyses the second examination serves as the “baseline” time for the cholesterol measurement.

Approximately 55% of those who entered the Framingham study were male; their mean age was 44 (median 43; range 28—62 years). By the time of the tenth examination, more than a third of the respondents were 65 years old or older. At the beginning of the study, none of the participants took antihypertensive medications. Presumably, in the ten examinations analyzed here, few received treatment for hypercholesterolemia.

For men and women combined, the mean cholesterol level was 228, and the median was 225 mg/dl. Table 1 compares the distributions of cholesterol levels among Framingham participants and among 20–74 year-old Americans enrolled in the most recent National Health and Nutrition Examination Survey. The Framingham values are slightly higher, reflecting perhaps the secular decline in cholesterol levels among Americans (National Center for Health Statistics, 1987) as well as differences between the Framingham population and the general U.S. public.

Q-Q plots like Figure 1a demonstrate that the marginal distribution of cholesterol at a particular exam is not Gaussian. A fortiori, the joint distribution across exams is not Gaussian. Figure 1b, a Q-Q plot using the logarithmic transformation of cholesterol, shows that the distribution fits well, except at the tails. In previous analyses we found that (raw) cholesterol is also consistent with a gamma distribution (Garber et al., 1990).

Dynamic variation in cholesterol levels can be appreciated in plots of cohort means
over time. Figure 2a shows the mean cholesterol level, along with the 5th, 25th, 75th, and 95th percentiles, for men 45 years of age and younger; Figure 2b graphs the same quantities for women. There are clear age/time trends in cholesterol levels, with a peak at exam 7 for both sexes, despite the substantial variability across individuals at a given exam.

These plots of cohort data provide no information about within-individual variability over time. Figure 3, which consists of plots of cholesterol levels over time for ten randomly selected men and women, shows that within-individual variability is substantial. Although the person whose exam 2 cholesterol level began at 350 mg/dl shows a clear downward trend, for many of the other observations no trend is apparent, and the levels fluctuate greatly. The plots indicate that the rankings of individuals’ cholesterol levels, particularly if the levels begin far apart, tend to be maintained.

The inter-examination correlation of cholesterol levels reflects these phenomena. Table 2 contains the correlation matrix for log cholesterol for women 45 years of age and younger; the matrix for men is similar. As one would expect, initially inter-examination correlation diminishes as the examinations are more widely separated. However, at least a component of the correlation persists. For example, the correlation between adjacent exams is closest for exams 6 and 7, at 0.826, yet even for examinations 2 and 10, which are 16 years apart, the correlation is 0.586. Thus, the cholesterol level demonstrates moderate persistence over time.

Because our analyses are based on very long-term followup, missing data is a potentially significant issue. Study participants occasionally missed blood tests, so at some exams their cholesterol levels were not measured. But more importantly, attrition due to death or illness could be important and is not likely to have been random. Because individuals with very low or very high cholesterol levels have excess mortality (Martin et al., 1986), the time path of an individual who was present for the entire 20 years may have been quite different from that for an individual who died before exam 5. For the analyses we report here, we used only individuals who had not died before the end of the period of observation. We determined how sensitive our results were to the choice of this nonrandom sample of Framingham participants by plotting cholesterol levels by age for the entire sample, including those for whom data were missing, and then plotting the same
quantities for individuals for whom complete data were available. Cholesterol levels over time were virtually identical for the two samples, despite the significant dropout rate in the later years of the study. Most of the analyses that follow are based on data from 708 men and 735 women.

III. ARE CHOLESTEROL LEVELS MARKOVIAN?

Perhaps the most attractive and natural approach to predicting changes in cholesterol would be to treat the time path of cholesterol levels as a Markov process, where the state of the system is the cholesterol level. In addition to providing direct estimates of transition probabilities, a Markov model, if appropriate, would make possible simple and elegant descriptions of the time course of measured cholesterol levels and enable us to find the probabilities of developing a high-risk cholesterol level at various times in the future.

To test the Markov model, we considered three nested hypotheses. The most restrictive hypothesis is that cholesterol is a stationary and Markovian stochastic process (so that its univariate marginal distributions are invariant with respect to its constant transition mechanism). The next most restrictive hypothesis is that cholesterol levels are Markovian with stationary transition mechanism, while the least restrictive alternative posits that they are Markovian with a nonstationary transition mechanism. The most restrictive hypothesis can be rejected on the basis of the simple observation that mean cholesterol values vary significantly with age; this is suggested by cross-sectional data and validated by our longitudinal data. The latter two hypotheses are not so readily rejected. We next examine the least restrictive hypothesis, since if it is inconsistent with the data, then any more restrictive alternative must also be rejected.

We divided the data into demographic subgroups, based on age and gender (race was not used as a variable, since the Framingham Heart Study includes very few nonwhite participants). Although the findings reported here are based on men and women who were 45 years of age or younger at entry to the study, preliminary analyses indicated that our most important conclusions hold regardless of age at entry. We also divided cholesterol levels into five discrete categories, which we retained for all of our analyses: $< 161 \text{ mg/dl;}$ $161 - 200 \text{ mg/dl;}$ $201 - 240 \text{ mg/dl;}$ $241 - 280 \text{ mg/dl;}$ and $> 280 \text{ mg/dl.}$ Three of the boundaries for these categories are thresholds for changes in management according
to the clinical guidelines of the (U.S.) National Cholesterol Education Program (NCEP). The NCEP labels a cholesterol level below 200 as "desirable"; 201–240 is considered to be moderately elevated ("borderline high risk"), and 241 and above is considered to be a high-risk cholesterol level. For men and women between the ages of 20 and 74, the cutpoint of 160 is approximately the 15th percentile; 200 is roughly the median; 240 is near the 75th percentile; and 280 is near the 90th percentile. The overriding consideration in defining these ranges, however, was that the changes from one state to another might lead to a change in patient management.

Three techniques were used to estimate the transition probabilities, which are displayed for women age 45 and younger in Table 3. Table 3a contains the empirical transitions from cholesterol categories during exam 2 to the categories at exam 3. For the transition probabilities in Table 3b we assumed that the cholesterol levels in exams 2 and 3 have a joint normal distribution whose means and covariance matrix are estimated from the data. We then used the joint distribution to calculate the conditional (transition) probability that the exam 3 cholesterol level will be in a particular range, given a range of exam 2 cholesterol levels. In Table 3c we follow the same procedure after applying a logarithmic transformation to the cholesterol levels. As the table makes clear, the log-transformed data using the joint normal assumption produce transition probabilities that are close to the empirical transition probabilities. With only a few exceptions, the probabilities in 3c are closer than those in 3b to the empirical probabilities, often substantially so. This provides further support for the use of log (cholesterol) in subsequent analyses. For our tests of the Markov property, we use the empirical transition probabilities.

Our method for testing the consistency of the data with a Markovian assumption is based on the Chapman-Kolmogorov equations. If for three successive time periods, say 1, 2, and 3, it is not the case that $P(1,2)P(2,3)=P(1,3)$ (where the $P$'s are estimated transition matrices for the indicated intervals of time) to within noise, then the data could not be Markovian. We compute chi-square statistics based, for example, on the $(1,2)$ and $(2,3)$ transitions to compute expected numbers of transitions for comparison to the observed $(1,3)$ transitions. To be precise, we calculate

$$\sum_{(k,r):n_{k,r}\geq 1}(n_{k,r}\hat{P}_{k,r}(1,3))^{-1}(n_{k,r}(\sum_{j} \hat{P}_{k,j}(1,2)\hat{P}_{j,r}(2,3) - \hat{P}_{k,r}(1,3)))^2,$$
where the karets indicate maximum likelihood estimates of parameters. Attained significance levels for the Markovian null hypothesis are computed by a bootstrap technique (Efron, 1982) that we devised for the purpose.

The validity of the technique rests in part on this simple observation. If \( \{X_m(t)\}_{m=1}^n \) are independent, finite Markov chains whose transitions are governed by transition matrices \( P(i, i+1) \), for \( i = 1, 2, \ldots \), (with \( k, r \) entries \( P_k, r(i, i + 1) \)) then given the events \( [X_m(i) = a(m)] \), \( m = 1, \ldots, n \), the random variables \( X_m(i + 1), m = 1, \ldots, n \) are independent; \( X_m(i + 1) \) is distributed as a single multinomial observation with parameter \( P_{a(m), r}(i, i + 1) \). We test whether the two-step (estimated) transition probabilities and the observed transition probabilities for the same interval are equal (within noise). Our approach differs from Anderson and Goodman’s (1957) test of the Markov property, which uses a likelihood ratio test to compare the suitability of one and two step fitted transition probabilities. Their work, and later comments by Hoeffding (1965, section 9), imply that the likelihood ratio test does not differ (to terms of order \( N^{-1/2} \)) from a corresponding chi-square test, at least if the probability of Type I error is constant or tends to 0 sufficiently slowly with increasing sample size. The fundamental difference between our test and theirs is the alternative hypothesis. Anderson and Goodman test the first-order Markov property against the specific alternative of second-order behavior. We directly test the validity of the Chapman-Kolmogorov equations, rather than compare first and second order processes. Even though our test may not be as powerful as the Anderson-Goodman test for the alternative that they consider, we show that it is very powerful in the context of our data and models. Furthermore, unlike their test, our bootstrap criterion for judging attained significance is not based on asymptotic distribution theory.

To test the Markov property, we used the empirical \( P(1, 2) \) and \( P(2, 3) \) matrices and multinomial simulations to generate pseudo-random sample paths from a process known to be Markovian with transitions governed by the estimated transition matrices; we infer from our simulated data the sampling distribution of the chi-square statistic under the null Markovian hypothesis. The bootstrap technique is a substitute for exact sampling distributions, which we do not know, and for theory that is asymptotic in cohort size or time, whose validity in this situation is questionable. More details of the simulations are given by pseudo-code in Appendix I. The results are displayed in Figure 4a, which is typical of other age-sex groups. The 99th percentile of the chi-square statistic is less than 30 if
the Markovian hypothesis is true. In contrast, the computed value for this cohort is 89.6, convincingly rejecting the Markovian hypothesis when three consecutive time periods are considered. Therefore, it cannot be supported by the data when more time periods are considered.

Although Type II error is not an issue here, we also assessed the power of our bootstrap-based test of the Markovian null hypothesis. We simulated data from a process that had the same one step conditional probabilities as were observed for the actual data, but which was constructed so as not to be Markovian. Indeed, the true conditional probability \( P(\text{Cholesterol in state } k \text{ at time 3} \mid \text{state } j \text{ at time 2 and state } i \text{ at time 1}) \) was deliberately constructed so as to depend on \( i \), though the true \( P(\text{Cholesterol in state } k \text{ at time 3} \mid \text{state } j \text{ at time 2}) \) was taken from observed data for all \((j, k)\). As is evident from Figure 4b, this bootstrap-based test has substantial power against alternatives that are germane to our modeling problem.

Had the discrete Markovian assumption been validated, then we could have investigated its validity in continuous time by utilizing a second approach, as found in Johansen (1974), Kingman (1975), and Frydman and Singer (1979). Since cholesterol is not Markovian in discrete time, it cannot be Markovian in continuous time, as it should be if it is Markovian at all.

The Markov property could have failed because we discretized the (continuous) cholesterol levels, because the cholesterol levels were measured with error, or because of "heterogeneity." To check the effect of discretization, we simulated as many (nonstationary) autoregressive sample paths with Gaussian innovations as there were in our cohort data. The simulated data were, of course, from a continuous state Markov process; they had initial distribution identical to the empirical initial distribution of measured cholesterol; and means, variances, and covariances of these simulated data were identical to those of the data that we used for testing (the marginal distributions were quite similar). For each sample path we grouped cholesterol values into the five discrete categories described above and computed the chi-square statistic for the Markov hypothesis as before. The empirical distribution of these statistics is nearly identical to the upper graph in Figure 4. Although discretization can destroy the Markov property, it is not the reason that our data fail to appear Markovian.
Errors in measurement offer another potential explanation, since adding independent random quantities to otherwise Markovian data can also destroy the Markov property. There is noise in the measurement of cholesterol levels, but intra-laboratory variation in cholesterol measurements has a coefficient of variation of only about 2-3% (Mogadam et al., 1990; Thompson and Pocock, 1990), a number far too small to explain the failure of the Markov property as judged by our test.

Higher-order state dependence can reflect “heterogeneity” due to inter-individual variation in the values of other risk factors. Analyses that condition on the values of the risk factors can often remove heterogeneity from this source. To assess the extent to which risk factors measured as part of the Framingham Study might be responsible for variation in rates of change in cholesterol levels, we tested regression models using a variety of covariates. Apart from age and sex, the covariates had weak and inconsistent associations with changes in cholesterol levels, and a first-order Markov process did not seem to fit the residuals from these regressions.

None of these factors seems sufficient to explain why the first-order Markov property did not hold. Higher order dependence is apparent in the correlations, covariances, and variograms (Diggle, 1990), which reveal non-negligible long run dependencies in the joint distributions within individuals. The variances of differences in cholesterol levels drawn at different examinations are smaller for adjacent examinations than for those that are distant in time, but the variances appear to level off as the intervals between exams grow. We tested a more general specification for these dependencies, an autoregressive-moving average process of order 1, or ARMA(1,1), and developed a novel technique to test its validity, in order to capture the near-term dependencies. Details are described in Appendix II. Our tests show that the ARMA(1,1) process fits the data very poorly, and although a very high-order ARMA process might be consistent with the data, no low-order ARMA process is likely to fit well. Because neither the Markov model nor this simple generalization were found to be suitable, we pursued alternative approaches to prediction.

IV. REGRESSION MODELS FOR CHOLESTEROL PREDICTION

Our specifications of alternative models to predict cholesterol levels were guided by consideration of the policies that we considered to be feasible because they are simple,
acceptable to patients and physicians, and of reasonable cost. We assume that cholesterol screening tests will be performed at approximately equal intervals, although the time between screening tests might be modified if for any reason an individual is found to be at high risk of subsequently developing an elevated cholesterol level. For example, a person who consistently has a cholesterol level that approaches the threshold for a "high" level might be tested sooner than one whose cholesterol level is consistently low.

Because our interest is in using cholesterol tests for screening, we remove from our analyses those persons whose cholesterol levels are considered high, after the time at which the high cholesterol would be discovered. For example, if 240 mg/dl is the cutoff for high cholesterol, and we consider policies that involve testing every two years, a person whose cholesterol level is 265 mg/dl at exam 3 is removed from the test population at that time. Such individuals would ordinarily receive additional cholesterol tests to assess the need for, and monitor response to, treatment; this use of cholesterol testing is outside the scope of our analysis. When we consider policies that include testing only every other exam (that is, every four years), we might not test at exam 3, so the earliest that the high cholesterol level would be discovered is at exam 4. If the measured cholesterol level at exam 4 was less than 240, a high cholesterol level would not be detected at either exam, and thus the person would not be censored, even then.

Usual approaches to prediction rest on the implicit assumption that the current cholesterol level is the same as the most recently measured level. The "predicted" cholesterol level is a constant, changing only when a new cholesterol level is obtained. We now investigate alternative approaches that use more information, such as previous cholesterol levels obtained under the screening policy, to predict future values, and to account for such factors as the upward (followed by downward) drift in cholesterol levels that accompanies aging. Approaches that use more information are expected to improve prediction, if only because multiple cholesterol measurements reduce the influence of transient cholesterol variability.

Regression models offer a simple and convenient approach to prediction. Because our Q-Q plots demonstrate that log cholesterol, but not untransformed cholesterol, is approximately Gaussian, we considered a variety of regression specifications, including Box-Cox transformations as well as logarithmic transformations of both independent and dependent
variables. We also investigated biased regression techniques (ridge regression with ridge parameter selected by cross-validation) in order to assess whether such approaches would reduce the mean square error of prediction.

In the scale of our measurements, cholesterol changes smoothly over time. So whatever its "true" probability structure might be, "best" prediction for an individual at times subsequent to C(8), for example, based on C(8) and several prior measured values, can only be a smooth function of these measured values and time. The Stone-Weierstrass theorem (Dunford and Schwartz, 1958, p. 272) implies that the predictor must be approximately a polynomial in these quantities. So various such polynomial functions were fit and validated. The simplest functions fit best. These were either affine in the previous cholesterol values or polynomials in time. Splines (local polynomials with smoothness constraints across their "knots") competed successfully with more conventional polynomials.

Our first regression approach treated the exam 9 cholesterol level as the dependent variable, with a set of previous cholesterol levels as the predictors. In this approach, time is not explicitly used for prediction, as would be proper if measured cholesterol were jointly Gaussian. Our sets of predictors were cholesterol levels obtained at exams 2, 4, 6, and 8 (corresponding to a policy of screening every four years); exams 2, 5, and 8 (every six-year screening policy); and exams 2 and 6 (eight-year screening policy).

To compare the fit of the alternative models, we used five-fold cross validation, implemented as follows. Suppose that $f_1$ and $f_2$ are transformations for the predictor and dependent cholesterol levels, respectively. Suppose that the predictors are the cholesterol levels at exams 2, 4, 6, and 8, and that the dependent variable is the cholesterol level at exam 9. We first randomly divide the cohort into five groups. We then leave one group out, regress $f_2(9)$ on $f_1(2,4,6,8)$, and predict $f_2(9)$ for the group left out from the regression equation. We then compute the residual sum of squares from the regression, where the residuals are expressed in terms of the original cholesterol scale, i.e., $C(9)$; we repeat this five times, leaving out each group once, then sum the squared residuals from each of these repetitions and average over all subjects. Five-fold cross-validation has been found by Breiman and Spector (1990) to be at least as good as competing methods for variable selection in regression.

This five-fold cross-validation scheme demonstrated that the residual sum of squares
is minimized when $f_1$ and $f_2$ are both in log scales. We rejected the Box-Cox and ridge regression approaches on this basis. Despite the long-run dependencies in measured cholesterol, the value at the most recent exam is the most influential in predicting the value at an exam; for example, exam 8 is most influential for predicting the exam 9 cholesterol level. Exclusion of this most recent value also increases the residual sum of squares substantially.

We found that the best approach for predicting exam 9 while using four prior examinations was to include exams 2, 4, 6, and 8 as predictors. With cross-validation, predictions based on several examinations need not fit better than more parsimonious specifications. For example, for women age 45 and under, models that use exams 2, 5, and 8 predict exam 9 better than those using 2, 4, 6, and 8. Results of the regressions predicting exam 9, using the log transformation of cholesterol and the five-fold cross-validation, appear in Table 4.

As an alternative to predictions based on linear regression, we considered spline regression techniques based on logged cholesterol values. Perhaps the most obvious advantage of a spline modeling approach is that it explicitly incorporates time (or age) as a predictor (and thus accommodates the joint distributions of logged cholesterol that are not exactly Gaussian). In addition, because they generate predictors that are (locally) polynomial functions of time, spline models provide prediction that is continuous and not restricted to a fixed set of examination times. This feature permits us to address a broader set of policy-determined intervals between screening tests than can be accommodated by a simple conventional regression in which (the log of) cholesterol at a particular future examination is predicted from (the logs of) cholesterol levels at a preassigned set of past times.

We tested a variety of spline specifications in the spirit of Rice and Silverman (1991). The analyses reported here use a particularly simple set of spline specifications and do well in terms of prediction. Consider an individual who has his cholesterol measured at exams 2, 4, 6, and 8. We seek to predict his cholesterol level at any subsequent time. We use a learning sample to estimate the functional relationship between the exam 10 cholesterol level and the values at exams 2, 4, 6, and 8. For a person in the test sample, we use this equation and his or her actual cholesterol levels at exams 2, 4, 6, and 8 to obtain a predicted exam 10 level. We then fit a quadratic spline with a knot at exam 5 and a constraint that the fitted value of the cholesterol level should equal the estimated cholesterol at exam 10. We compared this scheme to ordinary linear regression, as well as to pure quadratic and
cubic splines that were not constrained to equal the estimated cholesterol at exam 10. The linear regression uses information from examination 9 in the learning sample and can only predict the exam 9 value for the new person. Five-fold cross-validation demonstrated that a quadratic spline with one knot and the cited constraint appeared to fit the cholesterol level at exam 9 best.

Thus, our prediction methods emphasize linear fits that are affine in (the logarithms of) previous values of cholesterol, and spline regressions that explicitly incorporate time. Both approaches demonstrate that the use of prior cholesterol levels can offer substantial information about the likelihood that a person will have a high cholesterol level in the future. However, using the residual sum of squares as a criterion for fit is not an ideal measure of the model's utility for policy purposes. In fact, a model that fits less well in terms of the residual sum of squares might actually be better for policy purposes, as long as it infrequently results in misclassification. It is not crucial, from a policy point of view, to distinguish between cholesterol levels of 280 and 400 mg/dl, because both values are high enough to lead to additional testing or treatment. An error of 120 mg/dl, however, would contribute substantially to the residual sum of squares. Consequently, we turn to ROC-curve analysis in order to assess model fit and to compare the alternative prediction methods from the point of view of their usefulness for clinical and policy purposes.

V. MEASURING PREDICTIVE ACCURACY

In using cholesterol levels for clinical purposes, errors have consequences only if they lead to changes in management. These considerations led us to seek measures of predictive accuracy that penalize misclassification. ROC curves are a popular method for displaying misclassification rates, in which the true positive rate for the diagnostic or other test is plotted as a function of the false positive rate (Swetts and Pickett, 1982; McNeil and Hanley, 1984; Hanley, 1989).

We use the following procedure to draw the ROC curves. Denote by $\hat{C}_i(t)$ the predicted period $t$ cholesterol level for individual $i$. The prediction can be conditioned upon prior cholesterol levels and on the values of several covariates. Denote by $\hat{C}_i^+(t)$ the threshold predicted cholesterol level at time $t$. The threshold level has the following interpretation: a predicted cholesterol level that equals or exceeds the threshold is considered to be a
"positive" or abnormal value, while a predicted value less than the threshold is considered to be "negative" or normal. For our purposes, the "gold standard" test that defines abnormality is the actual cholesterol level at time \( t \). Based on the cutoff of 240 mg/dl, the level that the National Cholesterol Education Program considers to be undesirably high, a "true positive" result is obtained when \( \hat{C}_i(t) \geq \hat{C}^\dagger(t) \) and \( C_i(t) \geq 240 \) mg/dl. Sensitivity and specificity can be calculated from the 2×2 table generated by the "test" results, the actual cholesterol level, and the value selected for \( \hat{C}^\dagger(t) \), corresponding to a single point on the ROC curve. The entire curve is obtained by varying the value of \( \hat{C}^\dagger(t) \). As \( \hat{C}^\dagger \) is reduced, both the true-positive and false-positive rates increase. We plotted ROC curves corresponding to different models for the prediction of future cholesterol levels in order to compare the performance of these alternative predictors.

Our approach to classification differs from standard classification techniques, such as linear discriminant analysis and logistic regression. The two "classes," in this context, are \( C(9) \geq 240 \) and \( C(9) < 240 \). Our approach is motivated by the specific features of cholesterol levels; in this context, the assumptions underlying the traditional discrimination techniques are not satisfied. The usual motivation of the Fisher linear discriminant is that there are two possible sampling distributions governing predictors, one per class, both Gaussian; the means of these two distributions should be different, but the covariances about the same. In particular, the sampling distributions of covariates, conditional on class membership, should be roughly Gaussian. For us, the predictors are \( C(2), C(4), C(6), \) and \( C(8) \); and, at least in the log scale, their unconditional sampling distribution is roughly Gaussian. The conditional distributions given class membership (determined by the 240 dividing line for \( C(9) \)) are not Gaussian at all. They differ in covariance structure and in other respects. Further, neither conforms readily to the exponential family assumptions that motivate logistic regression. Thus we chose the alternative approach of predicting the exact value of the exam 9 cholesterol level, and varying cutoffs for the predicted value to generate the ROC curve.

Global measures of test performance can be derived from the ROC curve. For example, the area under an ROC curve is often used as a unidimensional summary of the ability of a given test to discriminate between normals and abnormals. However, because programs for the detection and management of hypercholesterolemia typically require further confirmatory testing after an abnormal value, the cost associated with a false-negative
rate may be higher than the cost for a false positive. Consequently, only the range of the ROC curve that corresponds to a high true positive rate is clinically relevant. Rather than provide the summary measure of the area under the ROC curves, we therefore plot the curves so that their performance in the range of values deemed relevant by the reader can be compared.

In generating the ROC curves for exam 9 from predicted levels, we employed linear regression and spline models. The regression method uses the predicted (log) cholesterol from the regression of log cholesterol on past values of log cholesterol. Because the predictor is a functional of the empirical distribution of our learning sample cohort (as defined by sex and age at entry to the Framingham study), imputing learning sample true and false positive rates to subsequent subjects, based on the functionals, is subject to the potential bias described by Efron (1982, pages 5 and especially 52 et seq., and also 1983). To correct this bias, we adapted Efron's bootstrap technique to the data, as we describe in Appendix III. As expected, based on the large number of data points in our analysis, we found that the bias adjustment has negligible effects on our estimates. Bias correction is less germane to the spline approach to prediction than to the first approach because the original exam 9 value is not used in fitting the spline model.

Prediction of log $C(9)$ from prior cholesterol levels is part of each of these methods. Calculation of the TPR and FPR involves additional steps. One approach, which we call the Gaussian method, is based on the assumption that \{log $C(t)$\} has a joint Gaussian distribution. The mean and covariance matrix of the joint distribution of [log $C(2)$, ..., log $C(8)$, log $\hat{C}(9)$] are estimated from the raw and fitted values. Then for various values of the constant $\hat{C}^{\dagger}$,

$$P[\hat{C}(9) > \hat{C}^{\dagger}, C(9) > 240, C(8) < 240] \text{ and } P[C(9) > 240, C(8) < 240],$$

are calculated from a Gaussian distribution with estimated mean vector and covariance matrix. The ratio of the above two estimated probabilities is the true positive rate corresponding to the cutoff $\hat{C}^{\dagger}$. The second approach, which we call the nonparametric method, also estimates $C(9)$ as above. In the nonparametric method, however, the above probabilities are estimated empirically from the learning sample data.

Figures 5a and 5b display ROC curves for men and women, respectively, based on the nonparametric bootstrap estimation method, Gaussian estimation, the spline method,
and empirical prediction based on the exam 8 cholesterol level alone. It is obvious that the
different estimation methods do not produce dramatically different ROC curves, although
for men the ROC curves computed by other techniques nearly everywhere dominate the
empirical ROC curve. Note also that the ROC curves for men and women level out at
predicted cholesterol levels below about 205, meaning that using a lower cutoff increases
the false positive rate without increasing the true positive rate. Moving from a \( \hat{C}^\dagger \) of 200
to 150, for example, has virtually no effect on men’s true positive rate of about .95, but
increases the false positive rate from about .6 to nearly 1.0.

The results in Figure 5, with the exception of the exam 8 empirical method, are
all based on similar data. How much do the ROC curves change as information from
additional examinations is added to the prediction method? Figures 6a and 6b show ROC
curves corresponding to predictions based on linear regressions that use different numbers
of exams. The most striking finding concerns the remoteness of the examinations. For
both men and women, prediction using 2, 5, and 8 values generates ROC curves that
are nearly identical to predictions using exams 2, 4, 6 and 8, but loss of examination 8
information worsens predictive accuracy. In the case of men, prediction based on exams 2
and 6 provides very little information. The ROC curve based on exams 2 and 6 for women
is closer to the ROC curves based on the other methods, but here too, particularly in the
most interesting range of true positive rates between .7 and .9, prediction that does not
include exam 8 information is relatively inaccurate.

How different are the ROC curves based on the different methods? To help answer this
question, we developed a new bootstrap-based technique to describe and display confidence
sets for individual points on the curve. Apart from the obvious differences in assumptions
that underlie the “nonparametric” and “Gaussian” computations of points on the curves,
the algorithms for computing the respective confidence sets are nearly identical. The
difference is that the nonparametric curves are bias-adjusted, while the Gaussian curves
(based as they are on the assumption that the logarithms of measured cholesterol are
nearly jointly Gaussian) are not. In either case, for a fixed \( \hat{C}^\dagger \) we denote the center of a
candidate confidence set by \((FPR_0, TPR_0)\). Draw \( n^{(1)}_B \) bootstrap samples, each of size \(|\mathcal{L}|\)
from the learning sample \( \mathcal{L} \), where \(|\mathcal{L}|\) indicates the number of elements of \( \mathcal{L} \). For each,
compute an \((FPR^*, TPR^*)\) bootstrap ROC point. These \( n^{(1)}_B \) points are now our “raw”
data to be used in computing confidence sets. We compute many sets of a form suggested
by that of Hotelling's $T^2$. The $2\times2$ covariance matrix $\hat{\Sigma}_*$ to be used is computed from the $n_B^{(1)}$ bootstrap pairs. The confidence set is of the form

$$\left\{(FPR, TPR): (FPR - FPR_0, TPR - TPR_0) \hat{\Sigma}_*^{-1} (FPR - FPR_0, TPR - TPR_0) \leq k\right\}. \quad (1)$$

A further bootstrap process is used to infer confidence coefficients for different $k$. Thus, we sample with replacement $n_B^{(2)}$ points from the $n_B^{(1)} (FPR^*, TPR^*)$ points, and denote them by $(FPR^{**}, TPR^{**})$. Also, we compute $\hat{\Sigma}_{**}^{-1}$ from them. Then for various $k$ we compute

$$\#\left\{(FPR^{**}, TPR^{**}): (FPR^{**} - FPR_0, TPR^{**} - TPR_0) \hat{\Sigma}_{**}^{-1} (FPR^{**} - FPR_0, TPR^{**} - TPR_0) \leq k\right\}/n_B^{(2)}.$$

The process is repeated $n_B^{(3)}$ times. We average the empirical probabilities to compute $\hat{F}^*(k)$; note that the same simulated data are used at once for all $k$. Finally, we ascribe confidence coefficient $100(1 - \alpha)\%$ to (1) provided $\hat{F}^*-1(k) = 1 - \alpha$. For us, $n_B^{(1)} = 400; n_B^{(2)} = 400; n_B^{(3)} = 25$. Enormous savings in computation were realized by varying $\hat{C}^*$ (and therefore dealing simultaneously with the entire ROC curve) with each bootstrap sample.

The critical values and coverage probabilities of our ROC confidence regions are amenable to computations like those that apply to other bootstrap-$t$ procedures (Hall, 1988). Though we have not filled in all of the details, it is clear that under suitable regularity conditions our algorithm provides hyperaccurate estimates of both critical values and coverage probabilities.

The results of these analyses appear in Figure 7. The confidence regions are estimated either by the nonparametric bootstrap technique or by the technique that relies on Gaussian assumptions. The Gaussian confidence sets are much smaller than the nonparametric ones, reflecting the extreme differences in variability of bootstrap sampling distributions in the two cases. Note from the figures that the correlation matrices differ considerably by methodology. If the Gaussian assumption is valid, it greatly improves the ability to distinguish between different ROC curves. The bootstrap-based nonparametric confidence regions are more robust, but are not capable of distinguishing between ROC curves that are as close together as these. Under the Gaussian assumptions, the confidence regions are narrow enough to confirm that at the upper part of the ROC curve, corresponding to a false-positive rate of between 0.6-0.7, and a threshold predicted cholesterol level of about 200-205, the probability that an individual will develop a high-risk cholesterol level at exam 9 is very small.
VI. CONCLUSIONS

Longitudinal changes in cholesterol levels are of interest for both descriptive and predictive purposes. Our interest is primarily predictive. If future cholesterol levels could be predicted accurately, then the cholesterol level need not be measured frequently. There is a consensus that a five-year interval between screening tests is adequate to identify persons who develop high-risk cholesterol levels, a view that our results tend to confirm. But our analysis offers a method for determining who is likely to reach an "elevated" (that is, requiring further diagnostic testing or treatment) level before the next scheduled cholesterol measurement. Of course, the method could also be applied to related lipoproteins (such as HDL and LDL) with suitable data; it may be more important in that context because tests for measuring these lipoproteins are more expensive.

It is well known that an individual's cholesterol level fluctuates over time, and this has been offered as a reason that a single cholesterol measurement does not predict the risk of developing coronary heart disease as accurately as the average of multiple measurements. Even though the (linear) correlations between cholesterol levels drawn several years apart are unimpressive, our models (which are nonlinear in age) can predict future values with modest accuracy.

We have tried to develop the simplest models that can accurately predict cholesterol levels. They appear to have significant advantages over other methods. The simplest model we tested, the first-order Markov process, and the ARMA(1,1) model, were inconsistent with patterns observed in the data. Furthermore, the violations of the Markov property could not be attributed to discretization of the cholesterol levels, and are unlikely to be due to laboratory error in measuring cholesterol.

Empirical transition probabilities -- simple counts of the proportion of individuals who move from cholesterol category $i$ to category $j$ in a given number of years -- might predict reasonably well if applied to other populations. But this approach could only be used to predict for time intervals corresponding to the followup intervals in the Framingham Heart Study (every two years). Linear regression approaches are similarly limited. The spline-based prediction method provides continuous prediction, and provides predictive accuracy comparable to that of linear regression. The ROC curves provide a useful summary of the predictive accuracy of the alternative modeling approaches, measuring the probability of
misclassification rather than the goodness-of-fit.

Our approach to modeling is intended to facilitate prediction based on an individual's age, sex, and a limited number of (no more than four) previous cholesterol measurements. It does not make explicit the probability structure underlying cholesterol dynamics, the factors that cause cholesterol levels to fluctuate, or the differences in individual paths of cholesterol levels. By modeling the probability structure more explicitly, we have found (Garber et al., 1990) that there is substantial inter-individual variability in growth patterns of cholesterol levels. In that work, we treated the cholesterol level as a process which is the sum of four terms: a deterministic mean value function that reflects both aging and secular variation; an individual-specific time intercept; error of measurement; and a more complex fourth term. The fourth term, which is random, represents the dependence of current "true cholesterol" within an individual on its previous values. ARMA schemes may be flexible enough for use in modeling this fourth term, but may not be the best approach for clarifying the causes of variation in cholesterol.

Explicit modeling of the probability structure is helpful for understanding cholesterol dynamics, but because it requires substantially more cholesterol measurements than would be feasible in routine clinical circumstances, it is less suitable for our purposes here. The predictions of a full probability model might not differ greatly from those of our more empirical approach, and the model is certain to be a smooth function of previously available measurements and time, all that we have to use. Prevalent current approaches to model building indicate that it is typically best to overfit initially, and to prune back fits to those that validate well (Breiman et al., 1984, Buja et al., 1989, Friedman, 1991). We have employed such an approach, with 5-fold cross-validation, though not all candidate predictors are reported here.¹ However, those predictors that were parsimonious and that

¹ Our approach to cross-validation is expressly designed to validate the policy algorithm rather than a single predictor. By leaving out a fixed percentage of the data (A) rather than a single observation (B), we cross-validate the policy of predicting, say, C(9) affinely from C(2), C(4), C(6), and C(8), rather than the explicit predictor computed from our learning sample. So approach B validates conditional on the learning sample, while our approach A is unconditional with respect to the learning sample. This distinction is mentioned by Efron (1983) and is analogous to the distinction between Models 1 and 2 in the exposition of CART (Breiman et al., 1984, pp. 325–327).
validated well are reported and are the basis of our ROC curve computations.

The ROC curves for our predictive models do not differ greatly, as long as they include a recent cholesterol measurement. The ROC curves also indicate that, for individuals whose predicted cholesterol level is below about 200 mg/dl, the probability that their measured cholesterol level will be above 240 mg/dl in two years is less than 5 or 10%. For such individuals, waiting four years (rather than two years) for the next cholesterol measurement will infrequently lead to a delay in initiating treatment. For a condition like hypercholesterolemia, in which the benefits of treatment for many subgroups of people (particularly those without a history of coronary heart disease) are modest, the occasional delay in initiating treatment due to a false-negative test result may be acceptable. Of course, the “best” point to select on the ROC curve depends on the relative losses due to false-negative and false-positive results, but a cutoff value below about 190 will not reduce the false-negative rate substantially, and will increase the false-positive rate.

We believe that this approach, which applies to cholesterol measurements within the Framingham population, will prove useful for describing patterns of other cardiac risk factors and metabolic processes (such as the serum creatinine level) as well, and that the general approach will predict accurately in other populations.
ACKNOWLEDGEMENTS: We are grateful to Peter Bickel, Leo Breiman, Edmund Biden, Jerry Halpern, and Lincoln Moses for their advice. As always, none of them bears any responsibility for deficiencies that remain.

This work was supported by a Senior HSR&D Research Associate Award of the (U.S.) Department of Veterans Affairs, Grant HL46297 from the National Heart, Lung, and Blood Institute, Grant AG07651 from the National Institute on Aging, and by National Science Foundation Grant DMS 89-96526-03. Garber is a Henry J. Kaiser Family Foundation Faculty Scholar in General Internal Medicine.
REFERENCES


Laboratory Standardization Panel, National Cholesterol Education Program (1988). Cur-


Med., 54, 413–430.

Thompson, S.G., and Pocock, S.J. (1990). The variability of serum cholesterol bmeasure-

Toronto Working Group on Cholesterol Policy (1990). Asymptomatic hypercholestero-
APPENDIX I: Markov Pseudocode

Algorithm: Testing the Markov Assumption

Step 1. Input the cholesterol levels at the times at which it will be tested, say $C(t_1), C(t_2)$, and $C(t_3)$.

Step 2. Estimate the transition matrices $P(1,2)$, $P(2,3)$, and $P(1,3)$.

Simulation

Step 3. Simulate $\tilde{C}(t_2)$ from the marginal distribution of $C(t_1)$ with transition matrix $P(1,2)$; simulate $\tilde{C}(t_3)$ from $\tilde{C}(t_2)$ with transition matrix $P(2,3)$. Then $C(t_1), \tilde{C}(t_2), \tilde{C}(t_3)$ form a pseudo-realization of a Markov process with the same one step transition matrices as were estimated from $C(t_1), C(t_2), C(t_3)$.

Step 4. Get $\tilde{C}(t_3)$ by re-ordering $\tilde{C}(t_3)$ as follows: for each fixed state of $\tilde{C}(t_2)$, re-order its corresponding states of $\tilde{C}(t_3)$ so that $\tilde{C}(t_3)$ has the same state as $C(t_1)$ as frequently as is possible, subject to the constraint that $\tilde{C}(t_2) \rightarrow \tilde{C}(t_3)$ has the same ONE STEP transition matrix as $\tilde{C}(t_2) \rightarrow \tilde{C}(t_3)$.

Testing

Step 5. Note that $P(1,2) \cdot P(2,3) = P(1,3)$ holds if $C(t_1), C(t_2), C(t_3)$ are Markovian. Now treat $P(1,2) \cdot P(2,3)$ as the predicted value of $P(1,3)$, and compute a chi-square test statistic.

Step 6. Substitute $\tilde{C}(t_2), \tilde{C}(t_3)$ for $C(t_2), C(t_3)$ and repeat Steps 2 and 5.

Step 7. Substitute $\tilde{C}(t_3)$ for $\tilde{C}(t_3)$ and repeat Step 6.

Comparing

Step 8. From Step 5, we get the sample value of the chi-square statistic; Step 6 gives the simulated value of the chi-square statistic for the pseudo Markov process, and Step 7 yields that for the non-Markov process. The percentile of a chi-square statistic computed for the simulated two-step Markov data (in the empirical distribution of the statistic computed for the simulated Markov data) gives an estimate of the power of our test.
APPENDIX II: Autoregressive-moving Average Model

We used the following procedure to test the goodness of fit of the ARMA(1,1) model, analyzing data on 1156 women 45 years of age and younger. Let \( C(t) \) denote the cholesterol level at exam \( t \). We first removed the effects of age by obtaining residuals from a regression of \( C(t) \) on age at examination 2, denoted \( X \). For \( t=2,\ldots,10 \), and for every subject, we found the residual vector \( Z_t \) from the regression

\[
C(t) = A(t) + B(t)X + Z_t.
\]

We then estimated the parameters of the mean 0 ARMA(1,1) model, which we write in the notation of Box and Jenkins (1976, pp. 73–80)

\[
Z_t = \phi Z_{t-1} + a_t - \theta a_{t-1},
\]

whose autocovariance function \( \gamma_k \) satisfies the difference equation

\[
\gamma_k = \phi_1 \gamma_{k-1} - \theta_1 \gamma_{za}(k-1),
\]

where \( \gamma_{za}(k-1) \) is the cross covariance function between \( Z \) and \( a \),

\[
\gamma_{za}(k) = E[Z_{t-k}a_t].
\]

We estimated the three unknown parameters by the method of moments. Denote the covariance matrix of \( \{Z_t\} \) estimated under the ARMA(1,1) model as \( V_0 \), and denote the empirical (unconstrained) covariance matrix as \( V \). If the “null hypothesis” that ARMA (1,1) applies were true, then \( VV_0^{-1} \) would be the identity, and in particular have roots identically 1. This implies that \( tr(VV_0^{-1} - I)^2 \) differs from 0 only by noise. Again we assessed significance by a (parametric) bootstrap technique, repeatedly simulating as much data as were in our cohort from a Gaussian ARMA (1,1) process that was guaranteed to have true parameters as were estimated from the cohort data. With each simulation we computed \( V_0, V \), the roots, and the cited sum of squares. The empirical distribution of the sums of squares was then the sampling distribution by which significance of the test statistic was judged. The ARMA (1,1) hypothesis was rejected, and so decisively (\( P < .0005 \)) that it could not be explained by our failure to first take logarithms. Thus it is evident that our data do not conform to a low-order ARMA scheme.
APPENDIX III: Bootstrap-based Bias Adjustment

By and large our notation conforms to that of Efron (1983). Write $\mathcal{L} = \{x_1, x_2, \ldots\}$ for our learning sample cohort (say, of women age not more than 45 on entry to the study). Write $C_i(9)$ for $x'_i$’s measured cholesterol at exam 9, and $\hat{C}_i(9)$, for the predicted value (based on $C_i(2), C_i(4), C_i(6), C_i(8)$ and the regression equation computed from $\mathcal{L}$). We say $x_i \in \mathcal{L}_{TPR} \subset \mathcal{L}$ if $x_i \in \mathcal{L}$, $C_i(9) \geq 240$, and $C_i(8) < 240$. So $\mathcal{L}_{TPR}$ consists of those subjects for whom (i) prediction from the regression bears directly on whether $C(9)$ should be measured ($C(8) < 240$), and (ii) $C(9)$ indicates placement in the “high risk” group. We note as an aside that for both women and men whose ages did not exceed 45 on entry to the study, $\mathcal{L}$ had more than 700 subjects and $\mathcal{L}_{TPR}$ more than 300. In what follows the cardinality of $\mathcal{L}_{TPR}$ is written $n_{TPR}$; that is, $|\mathcal{L}_{TPR}| = n_{TPR}$.

Write $y_i = 1$ if $x_i \in \mathcal{L}_{TPR}$; otherwise $y_i = 0$. For $x_i \in \mathcal{L}_{TPR}$, and a candidate $c$ for which we might recommend observing $C_i(9)$ when $\hat{C}_i(9) \geq c$, write $\eta(x_i, \mathcal{L}) = 1$ if $\hat{C}_i(9) \geq c$; otherwise $\eta(x_i) = 0$. Write $Q[y_i, \eta(x_i, \mathcal{L})] = 1$ if $\eta_i \neq y_i$; otherwise $Q[y_i, \eta(x_i, \mathcal{L})] = 0$. So $Q$ counts “mistakes.” Then the apparent true positive rate is written

$$\overline{err} = n_{TPR}^{-1} \sum_{\mathcal{L}_{TPR}} Q[Y_i, \eta(x_i, \mathcal{L})].$$

The true error rate is written $Err$, and is $E\{Q[Y, \eta(X)] | \mathcal{L}\}$. Thus $Err$ is the conditional expectation, given the learning sample cohort, of $Q$ computed for a new subject. The understanding is that the new patient’s $X$ and $Y$ are distributed as those in (the iid) learning sample. The optimism, $op$, of the apparent true positive rate is $op = Err - \overline{err}$, so $Err = \overline{err} + op$; $op$ is estimated, by $\omega$, from bootstrapping as follows.

There are $B$ bootstrap samples $\mathcal{L}^*_1, \ldots, \mathcal{L}^*_B$. Each $\mathcal{L}^*_b$ is generated by random sampling with replacement from $\mathcal{L}$ a sample of size $|\mathcal{L}|$. Each $\mathcal{L}^*_b$ entails an $\mathcal{L}^*_{\{b, TPR\}}$ and $\eta$. For each $b$, $\hat{\omega}_b = n_{TPR}^{-1} \sum_{\mathcal{L}_{TPR}} Q[Y_i, \eta(X_i, \mathcal{L}^*_b)] - n_{TPR}^{-1} \sum_{\mathcal{L}_{TPR}} Q[y_i, \eta(x_i, \mathcal{L}^*_b)]$. Then $\hat{\omega} = B^{-1} \sum_{b=1}^{B} \hat{\omega}_b$. Note that $\hat{\omega}_b$ is motivated by the general bootstrap heuristic that the bootstrap sample bears the same relationship to the learning sample as the learning sample does to nature. That being the case, $\omega$ is a natural estimate of the optimism in the apparent error rate. We took $B = 10$. The resulting bias adjustment $\omega$ was not more than 1% of the TPR (and similarly for FPR), so we were not motivated to try refined estimators of $\omega$, such as the .632 rule (Efron, 1983).
Table 1. Comparison of cholesterol levels
Framingham Heart Study and 1976-80 NHANES

<table>
<thead>
<tr>
<th>AGE RANGE</th>
<th>AGEC</th>
<th>Mean</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men-NHANES (all races)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35–44</td>
<td>1</td>
<td>217</td>
<td>187</td>
<td>215</td>
<td>244</td>
</tr>
<tr>
<td>45–54</td>
<td>2</td>
<td>227</td>
<td>197</td>
<td>223</td>
<td>255</td>
</tr>
<tr>
<td>55–64</td>
<td>3</td>
<td>229</td>
<td>198</td>
<td>225</td>
<td>254</td>
</tr>
<tr>
<td><strong>Men-Framingham Exam 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35–44</td>
<td>1</td>
<td>226</td>
<td>198</td>
<td>221</td>
<td>249</td>
</tr>
<tr>
<td>45–54</td>
<td>2</td>
<td>230</td>
<td>205</td>
<td>228</td>
<td>253</td>
</tr>
<tr>
<td>55–64</td>
<td>3</td>
<td>229</td>
<td>198</td>
<td>226</td>
<td>256</td>
</tr>
<tr>
<td><strong>Women-NHANES (all races)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35–44</td>
<td>1</td>
<td>207</td>
<td>177</td>
<td>202</td>
<td>231</td>
</tr>
<tr>
<td>45–54</td>
<td>2</td>
<td>232</td>
<td>199</td>
<td>228</td>
<td>257</td>
</tr>
<tr>
<td>55–64</td>
<td>3</td>
<td>249</td>
<td>215</td>
<td>242</td>
<td>277</td>
</tr>
<tr>
<td><strong>Women-Framingham Exam 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35–44</td>
<td>1</td>
<td>212</td>
<td>184</td>
<td>205</td>
<td>234</td>
</tr>
<tr>
<td>45–54</td>
<td>2</td>
<td>240</td>
<td>209</td>
<td>234</td>
<td>267</td>
</tr>
<tr>
<td>55–64</td>
<td>3</td>
<td>256</td>
<td>225</td>
<td>250</td>
<td>287</td>
</tr>
</tbody>
</table>
Table 2. Correlation Matrix of Log CHOL for 735 Women Age ≤ 45

<table>
<thead>
<tr>
<th>Exam No.</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.000</td>
<td>0.731</td>
<td>0.735</td>
<td>0.706</td>
<td>0.667</td>
<td>0.650</td>
<td>0.649</td>
<td>0.611</td>
<td>0.586</td>
</tr>
<tr>
<td>3</td>
<td>0.731</td>
<td>1.000</td>
<td>0.768</td>
<td>0.764</td>
<td>0.718</td>
<td>0.689</td>
<td>0.662</td>
<td>0.641</td>
<td>0.597</td>
</tr>
<tr>
<td>4</td>
<td>0.735</td>
<td>0.768</td>
<td>1.000</td>
<td>0.811</td>
<td>0.797</td>
<td>0.748</td>
<td>0.711</td>
<td>0.661</td>
<td>0.632</td>
</tr>
<tr>
<td>5</td>
<td>0.706</td>
<td>0.764</td>
<td>0.811</td>
<td>1.000</td>
<td>0.821</td>
<td>0.768</td>
<td>0.752</td>
<td>0.712</td>
<td>0.666</td>
</tr>
<tr>
<td>6</td>
<td>0.667</td>
<td>0.718</td>
<td>0.797</td>
<td>0.821</td>
<td>1.000</td>
<td>0.826</td>
<td>0.769</td>
<td>0.721</td>
<td>0.650</td>
</tr>
<tr>
<td>7</td>
<td>0.650</td>
<td>0.689</td>
<td>0.748</td>
<td>0.768</td>
<td>0.826</td>
<td>1.000</td>
<td>0.804</td>
<td>0.756</td>
<td>0.684</td>
</tr>
<tr>
<td>8</td>
<td>0.649</td>
<td>0.662</td>
<td>0.711</td>
<td>0.752</td>
<td>0.769</td>
<td>0.804</td>
<td>1.000</td>
<td>0.803</td>
<td>0.729</td>
</tr>
<tr>
<td>9</td>
<td>0.611</td>
<td>0.641</td>
<td>0.661</td>
<td>0.712</td>
<td>0.721</td>
<td>0.756</td>
<td>0.803</td>
<td>1.000</td>
<td>0.775</td>
</tr>
<tr>
<td>10</td>
<td>0.586</td>
<td>0.597</td>
<td>0.632</td>
<td>0.666</td>
<td>0.650</td>
<td>0.684</td>
<td>0.729</td>
<td>0.775</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 3. Transition Matrix of Exam 3 Given Exam 2 for 735 Women Age ≤ 45

(a) Empirical Calculation

<table>
<thead>
<tr>
<th>category</th>
<th>(0, 160)</th>
<th>[160, 200)</th>
<th>[200, 240)</th>
<th>[240, 280)</th>
<th>[280, +∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 160)</td>
<td>0.404</td>
<td>0.532</td>
<td>0.064</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>[160, 200)</td>
<td>0.071</td>
<td>0.557</td>
<td>0.324</td>
<td>0.047</td>
<td>0.000</td>
</tr>
<tr>
<td>[200, 240)</td>
<td>0.022</td>
<td>0.223</td>
<td>0.446</td>
<td>0.259</td>
<td>0.050</td>
</tr>
<tr>
<td>[240, 280)</td>
<td>0.000</td>
<td>0.008</td>
<td>0.356</td>
<td>0.432</td>
<td>0.203</td>
</tr>
<tr>
<td>[280, +∞)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.128</td>
<td>0.410</td>
<td>0.462</td>
</tr>
</tbody>
</table>

(b) Gaussian Assumption for Cholesterol Levels

<table>
<thead>
<tr>
<th>category</th>
<th>(0, 160)</th>
<th>[160, 200)</th>
<th>[200, 240)</th>
<th>[240, 280)</th>
<th>[280, +∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 160)</td>
<td>0.458</td>
<td>0.435</td>
<td>0.102</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>[160, 200)</td>
<td>0.115</td>
<td>0.455</td>
<td>0.359</td>
<td>0.055</td>
<td>0.016</td>
</tr>
<tr>
<td>[200, 240)</td>
<td>0.014</td>
<td>0.181</td>
<td>0.439</td>
<td>0.211</td>
<td>0.155</td>
</tr>
<tr>
<td>[240, 280)</td>
<td>0.001</td>
<td>0.026</td>
<td>0.201</td>
<td>0.298</td>
<td>0.475</td>
</tr>
<tr>
<td>[280, +∞)</td>
<td>0.000</td>
<td>0.004</td>
<td>0.078</td>
<td>0.264</td>
<td>0.655</td>
</tr>
</tbody>
</table>

(c) Gaussian Assumption for Log Cholesterol Levels

<table>
<thead>
<tr>
<th>category</th>
<th>(0, 160)</th>
<th>[160, 200)</th>
<th>[200, 240)</th>
<th>[240, 280)</th>
<th>[280, +∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 160)</td>
<td>0.427</td>
<td>0.512</td>
<td>0.059</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>[160, 200)</td>
<td>0.063</td>
<td>0.541</td>
<td>0.345</td>
<td>0.048</td>
<td>0.004</td>
</tr>
<tr>
<td>[200, 240)</td>
<td>0.005</td>
<td>0.226</td>
<td>0.504</td>
<td>0.220</td>
<td>0.045</td>
</tr>
<tr>
<td>[240, 280)</td>
<td>0.000</td>
<td>0.048</td>
<td>0.340</td>
<td>0.407</td>
<td>0.204</td>
</tr>
<tr>
<td>[280, +∞)</td>
<td>0.000</td>
<td>0.007</td>
<td>0.128</td>
<td>0.368</td>
<td>0.497</td>
</tr>
</tbody>
</table>

33
### Table 4

5-fold cross validation:

<table>
<thead>
<tr>
<th>Predictor cholesterol levels</th>
<th>Residual sum of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exam(s)</strong></td>
<td></td>
</tr>
<tr>
<td>2,4,6,8</td>
<td>688.7497</td>
</tr>
<tr>
<td>2,5,8</td>
<td>719.6465</td>
</tr>
<tr>
<td>8</td>
<td>779.8674</td>
</tr>
<tr>
<td>7,8</td>
<td>670.2851*</td>
</tr>
<tr>
<td>6,7,8</td>
<td>654.9208*</td>
</tr>
<tr>
<td>2-7</td>
<td>726.5415</td>
</tr>
<tr>
<td>2-6,8</td>
<td>697.2442</td>
</tr>
<tr>
<td>2-5,7,8</td>
<td>648.4302</td>
</tr>
<tr>
<td>2-4,6-8</td>
<td>644.4892</td>
</tr>
<tr>
<td>2,3,5-8</td>
<td>646.9870</td>
</tr>
<tr>
<td>2,4-8</td>
<td>649.6454</td>
</tr>
<tr>
<td>3-8</td>
<td>647.8778</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictors cholesterol levels</th>
<th>Residual sum of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,4,6,8</td>
<td>765.9295</td>
</tr>
<tr>
<td>3,5,7,9</td>
<td>624.7736** (Exam 9 is crucial)</td>
</tr>
<tr>
<td>2,5,8</td>
<td>805.3395</td>
</tr>
<tr>
<td>3,6,9</td>
<td>657.7210**</td>
</tr>
<tr>
<td>9</td>
<td>702.0773</td>
</tr>
<tr>
<td>8,9</td>
<td>672.0231</td>
</tr>
<tr>
<td>2-9</td>
<td>640.4620 (big model is worse this time)</td>
</tr>
</tbody>
</table>
Figure 1. QQ normal plots for the cohort of 735 women 45 years of age and younger at exam 2. Panel (a): original scale of cholesterol levels; Panel (b): logarithm of cholesterol levels.
Figure 2 (a). Box-like plots for the cohort of 708 men 45 years of age and younger at exam 2. The solid line is obtained by linearly connecting the means of the cholesterol levels at each exam. The heavy dashes are 25th and 75th quantiles of the cholesterol levels at each exam, and the lighter dashes 5th and 95th quantiles.
Figure 2 (b). Box-like plots for the cohort of 735 women 45 years of age and younger at exam 2. The solid line is obtained by linearly connecting the means of the cholesterol levels at each exam. The heavy dashes are 25th and 25th quantiles of the cholesterol levels at each exam, and the lighter dashes 5th and 95th quantiles.
Figure 3. The paths of the cholesterol levels for all (1445) persons 45 years of age and younger at exam 2. The heavier solid line is obtained by linearly connecting the means of the cholesterol levels at each exam, and the dotted lines are the 25th and 75th percentiles. Other ten solid curves are obtained by linearly connecting the values of a random sample.
Figure 4. The power of testing Markovian assumption for exams 3, 4, and 5 in the group of 708 men age 45 and younger. The number of repetitions for bootstrapping is 300.
Figure 5. ROC curves using different estimation methods. Panel (a): 708 men 45 years of age and younger; Panel (b): 735 women 45 years of age and younger. bootstrap, ——; Gaussian, ······; spline, – – –; exam 8 (empirical), — —. Labels are thresholds for bootstrap curves.
Figure 6. ROC curves based on different sequences of exams. Panel (a): 708 men 45 years of age and younger; Panel (b): 735 women 45 years of age and younger. Exams 2, 4, 6, and 8, ——; exams 2, 5, and 8, ·····; exams 2 and 6, ——; exam 8 (empirical), ——. Labels are thresholds for the curves from exams 2, 4, 6, and 8.
Figure 7. Pointwise confidence regions for ROC curves. Panel (a): 708 men 45 years of age and younger. (b): 735 women 45 years of age and younger. Exams 2, 4, 6, and 8 are used to predict exam 9. (FPR, TPR) corresponding to the labeled threshold and obtained from empirical method, +; confidence regions centered at points on the empirical ROC curve, ·······; (FPR, TPR) obtained from Gaussian computation, *; confidence regions for Gaussian ROC curve – – –.