11-1-2002

Chronic Disease Data And Analysis: Current State Of the Field

Ralph D'Agostino Sr.
Boston University

Lisa M. Sullivan
Boston University

Chronic Disease Data And Analysis: Current State Of the Field

Ralph D'Agostino, Sr.
Boston University

Lisa M. Sullivan
Boston University

Chronic disease usually spans years of a person's lifetime and includes a disease free period, a preclinical, or latent period, where there are few overt signs of disease, a clinical period where the disease manifests and is eventually diagnosed, and a follow-up period where the disease might progress steadily or remain stable. It is often of interest to investigate the relationship between risk factors measured at a point in time (usually during the disease free or preclinical period), and the development of disease at some future point (e.g., 10 years later). We outline some popular designs for the identification of subjects and discuss issues in measurement of risk factors for analysis of chronic disease. We discuss some of the complexities in these analyses, including the time dependent nature of the risk factors and missing data issues. We then describe some popular statistical modeling techniques and outline the situations in which each is appropriate. We conclude with some speculation toward future development in the area of chronic disease data and analysis.

Keywords: Chronic disease, cardiovascular disease, Framingham Heart Study, logistic regression analysis, longitudinal data, missing data, mixed models, survival analysis

Introduction

A chronic disease is a disease first characterized by a development period or latent period in

Ralph B. D'Agostino, Sr., is Professor of Mathematics/Statistics, Public Health and Law. He is a fellow of the American Statistical Association and the Cardiovascular Epidemiology section of the American Heart Association. Email: ralph@bu.edu. Lisa M. Sullivan is Associate Professor of Biostatistics in the School of Public Health, Associate Professor of Mathematics and Statistics in the College of Arts and Sciences, and Associate Professor of Medicine in School of Medicine. She is Co-Director of the Graduate program in Biostatistics at Boston University. Email: lsull@bu.edu.

which the disease progresses subclinically. The latent period can be extensive in time. For example, in cardiovascular disease, build up of plaque in the arteries can begin in childhood. During this latent period the person often displays no overt effects or problems. Then the disease manifests itself in a clinical phase.

With cardiovascular disease, this may begin with a myocardial infarction (heart attack) where the heart suffers permanent injury due to the blockage caused by the plaque. After the appearance of the clinical phase, the affected person (or host) may follow a course that leads to little or substantial deterioration and possibly death.

In this example of cardiovascular disease, the clinical phase is initiated by a clinical event, a heart attack, and then followed
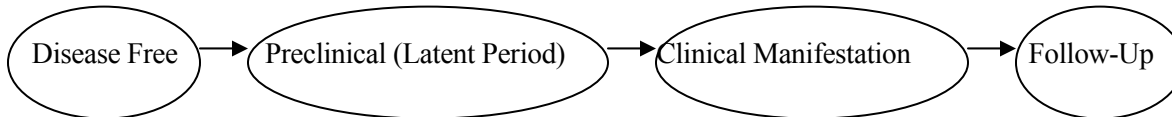
by a post event phase where there may be a general weakening of the body which increases the risk of subsequent cardiovascular events such as a second heart attack or a stroke resulting in death.

Lung cancer is an example of another chronic disease. Here the subclinical, latent period can consist of lung tumors developing over a period of more than 10 years before clinical manifestation and diagnosis. After diagnosis, there can be periods of stabilization, remission and progression. AIDS is still another example, where the subclinical stage can be characterized by a positive HIV infection. The clincal manifestation of AIDS

may then appear followed by a series of infections, increased deterioration and ultimately death. Alzheimer's disease provides an example where the distinction between the preclinical stage and clinical stage is blurred. In the preclinical phase, there is a progressive decline in cognitive function, especially noted in short term memory, and often personality changes. These ultimately lead to a stage where the person is unable to care for him or herself. The diagnosis of Alzheimer's disease often results when the person is debilitated and other forms of dementia (e.g., caused by a series of strokes) are ruled out.

A simple model for chronic disease is as follows: (1)

$$\text{Disease Free} \rightarrow \text{Preclinical (Latent Period)} \rightarrow \text{Clinical Manifestation} \rightarrow \text{Follow-Up}$$

Interest focuses on all four components. Each presents detailed and sophisticated modeling, data collection and analytic issues. Consider, for example, the 'Disease Free -> Preclinical (Latent Period) -> Clinical Manifestation'

component. This can be further refined to three submodels (shown below) where DF represents a completely disease-free state, PC represents preclinical signs and symptoms and C represents disease manifestation (clincal):

$$DF \longrightarrow PC \longrightarrow C \qquad\qquad (2.1)$$

$$DF \Big\langle \begin{array}{l} \nearrow PC1 \\ \searrow PC2 \longrightarrow C \end{array} \qquad\qquad (2.2)$$

$$DF \longleftrightarrow PC \longrightarrow C \qquad\qquad (2.3)$$

In (2.1), the disease free (DF) stage leads to the preclinical (PC) stage which in turn leads directly to the clinical stage (C). In such a situation knowledge of the preclinical stage could be useful in delaying or averting the clincal stage (C). Simple models of breast and colon cancer fit this situation. In (2.2), the disease free (DF) stage can lead to preclinical stages 1 or 2 (PC1 and PC2, respectively). PC1 does not progress to the clinical stage (C) while PC2 does. In this situation, identification of the

preclinical stage (PC) does not imply that the clincal stage (C) follows. Cervical cancer is an example of this situation. Lastly, (2.3) displays a situation where the preclinical stage (PC) may actually revert to the completely disease free (DF) stage or may lead to the clinical (C) stage.
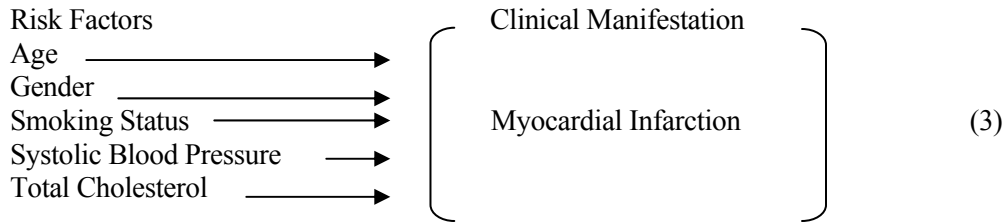
We could extend and elaborate the second component of model (1) 'Clinical Manifestation -> Follow-Up' in a similar fashion incorporating the complexities that are

involved in diagnosing the presence of the disease and the follow-up after that.

Chronic disease data and analysis questions relate to all aspects of the above (disease free, preclinical, clinical manifestation and follow-up). Good statistical approaches involve hypothesizing models for these aspects, collecting appropriate data, and then fitting and testing the appropriate models. Before fitting statistical models, biological models need to be formulated. Both (1) and (2) above represent simple models.

One important set of models relate risk factors (RF) of a disease free individual to the probability of manifestation of the clinical stage of the disease. For example, the relationship between age, gender, smoking status, blood pressure and cholesterol to the development of a myocardial infarction could be modelled as:

$$
\begin{array}{l}
\text{Risk Factors} \\
\text{Age} \\
\text{Gender} \\
\text{Smoking Status} \\
\text{Systolic Blood Pressure} \\
\text{Total Cholesterol}
\end{array}
\longrightarrow
\left\{
\begin{array}{l}
\text{Clinical Manifestation} \\
\\
\text{Myocardial Infarction}
\end{array}
\right\}
\qquad (3)
$$

To turn this into a statistical model one needs to decide how to identify appropriate (disease free) subjects, how many subjects to sample, when to measure the risk factors and how long to follow them. The latter item of follow-up is to ensure that a sufficient number develop a myocardial infarction so the components (or parameters) of the mathematical model can be estimated with good precision.

In a later part of this article we discuss in more detail the methods of statistical modeling for chronic disease. We discuss some popular designs for studies of chronic disease and we use cardiovascular disease as an example throughout the discussion. We review some of the methodologic issues that arise in studies of chronic disease and outline some popular statistical modeling and analysis techniques. We conclude with some speculation towards future developments. In the next section we present an example to motivate the discussion that follows.

2. Motivation: Cardiovascular Disease Example

Consider a study of cardiovascular disease, in particular a study of the risk factors associated with the development of cardiovascular disease. A first challenge is to understand the outcome, and in particular the conditions that should be considered part of the outcome and how they should be measured. A second challenge is to determine which risk factors should be measured and how frequently they should be measured in the study subjects. A related challenge is the specification of the appropriate statistical model to relate candidate risk factors to the outcome. In the following we illustrate the complexities of each step using cardiovascular disease as an example.

Defining the Outcome. Cardiovascular disease includes a number of conditions and is a major cause of morbidity and mortality worldwide. The most common serious cardiovascular disease is coronary heart disease (also called cardiac ischemia, defined as insufficient blood supply due to atherosclerosis of the coronary arteries). It consists of myocardial infarction (heart attack), which is direct damage to the heart, coronary deaths, and angina (persistent chest pain due to cardiac ischemia). Cardiovascular disease also includes other conditions such as stroke (or brain attack), and peripheral artery disease (circulation problems often in the calves). Cardiovascular disease is believed to have a long preclinical or latent stage.

For example, patients with coronary heart disease (CHD) are diagnosed (and enter the clinical stage) in a variety of ways. One patient may present with angina at an early

stage while another may suffer a heart attack after an otherwise asymptomatic history. Accurate determination of a cardiovascular event is critical, and the technologies to determine specific events are evolving over time. At one time an MI was mainly diagnosed by electocardiogram. Now it is standard to use enzyme tests (e.g., SGOT and CPK) Often chronic disease outcomes include condition-specific mortality (e.g., death due to cardiovascular disease). In such cases, elaborate protocols are required to ascertain outcome status. These include, in some cases, reviewing death certificates and/or hospital records. Determining cause of death can be further complicated by incomplete or ambiguous specification of the cause of death by the medical personnel evaluating the death.

Specifying the Risk Factors and the Data Collection Schedule. Determining the risk factors associated with the development of chronic disease (e.g., cardiovascular disease) requires an understanding of the biological complexity of the disease, some of which might change over time. Generally, studies of cardiovascular disease consider the following risk factors: gender, age, blood pressure, cholesterol, smoking status, and history of diabetes. Cardiovascular diseases span decades of individuals lives (from the preclinical to the clinical and follow-up stages).

Studies of cardiovascular disease often take years to complete, with the duration of the study influenced by the time it takes to observe a sufficient number of outcome events. The importance and influence of risk factors may vary over time (e.g., obesity at an early age and maintained over time may be important in leading to cardiovascular disease while the most recent blood pressure may be more important than blood pressure measured decades earlier). So, often risk factors are measured at the outset, and then repeated over the follow-up period. Investigators must decide what intervals are most appropriate to obtain repeat measurements. The interval is influenced by the stability (or lack of) of the risk factors over time.

For example, total cholesterol level is a relatively stable risk factor whereas smoking status is not. The latter would need to be measured on a more frequent basis. In recent studies of cardiovascular disease, investigators consider genetic and environmental factors, along with a broader array of clinical risk factors. In some cases, investigators have the flexibility to add new risk factors to a data collection protocol during an ongoing study. This introduces an analytic issue in that these new risk factors will not be measured on the same schedule as the core set (i.e. those measured since the outset). In cardiovascular disease, surgical procedures have also advanced rapidly in the last two decades and include introduction of artificial aortic valves, open heart surgery, angioplasty (opening blocked arteries using balloon catheters) and regulation of heart rythms by implanted pacemakers.

In parallel, pharmacologic treatments have become increasingly effective in treating known risk factors of cardiovascular disease (e.g., hypertension, hyperlipidemia) thereby slowing the manifestation and progression of disease. It is important to measure these interventions, which generally modify the effects of the risk factors on the development of disease, along with the risk factors themselves. Designs for studies of chronic disease and methlologic issues that arise in studies of chronic disease are discussed in detail in Section 3.

Choosing the Correct Model. The choice of the appropriate statistical model should be based primarily on a biological model. It should also be influenced by specific aspects of the design such as whether subjects are followed for a fixed period of time and then determined to have or not have the disease at the end of the observation period or whether subjects are followed for different amounts of time and have disease status ascertained at the end of the observation period. In a study of cardiovascular disease, a subject might die during the observation period due to cancer (or some disease other than cardiovascular disease) and at the time of death be free of cardiovascular disease. The most appropriate statistical model is one that utilizes all of the information that was measured on this person rather than exclude him or her because of the complexity of the data. Popular statistical models for studies of chronic disease are discussed in detail in Section 4.

3. Designs, Subject Selection and Data for Studies of Chronic Disease

The data for studies of the relationship between risk factors and development and progression of chronic disease can be prospective, retrospective or cross-sectional. Prospective study designs involve identifying individuals who are free of the disease of interest and following them over time. These studies can include repeated measurements of risk factors over time and monitoring for the development and progression of disease. The schedule for following individuals and repeating measurements depends on a number of factors including the stability of the risk factors over time and the nature of the relationship between the risk factors and disease status over time. Retrospective studies (also called case control studies) usually involve identifying two groups of individuals; those with the disease of interest (often called cases) and matches who are free of the disease of interest (often called controls).

Data are collected retrospectively usually by way of individual's recollection of prior health and risk behaviors or through medical record review. These studies are not optimal. It is usually difficult to assemble representative groups of cases and controls. Often the cases represent either the sickest (e.g., subjects enrolled through an Alzheimer's clinic) or the healthiest (e.g., those who have not died) of those affected with the disease. Further, the controls often differ in many ways from the cases, confounding the comparison of cases and controls. In addition, these studies can be subject to a number of biases (for example, recall bias or inaccurate recollection of specific behaviors or measurment based on incomplete medical records).

Cross-sectional studies are conducted at a point in time and represent concurrent risk factor and disease status. In some cross-sectional studies, individuals provide historical data on risk behaviors on the basis of recollection, thereby also subjecting these studies to recall bias.

Longitudinal cohort studies are most well suited for the analysis of chronic disease. We now describe in detail the specifics of longitudinal cohort studies and outline a well

known study of cardiovascular disease, the Framingham Heart Study.

3.1. Longitudinal Cohort Studies: The Framingham Heart Study

In longitudinal cohort studies, a group or cohort of individuals is assembled at the outset. The inclusion criteria often require a set of individuals to be free of the disease of interest. This is not always the case and those with prevalent disease may be enrolled at the outset. Individuals are followed prospectively in time. Serial measurements can be taken on a predetermined schedule, often at fixed time intervals (e.g., measurements every 2 years or every 5 years). Outcome or disease status is measured over time. For those individuals who develop disease, measures of the progression or severity of disease are also taken. There are several, large longitudinal cohort studies of cardiovascular disease, probably the best known study is the Framingham Heart Study, described below.

The Framingham Heart Study began in 1948 and is one of the most ambitious and daring longitudinal medical studies ever initiated. A cohort of 5,209 individuals, 2336 males and 2873 females, was enrolled from Framingham, MA. These represented a 60% sample of the town with ages from 28 to 62 years. Multiple risk factors were measured biennially, and the study continues today with surviving participants involved for over 50 years. Major cardiovascular risk factors have been measured since the outset (e.g., blood pressure, total cholesterol and smoking status) while others have been introduced as they were hypothesized to have an impact on the develoment of cardiovascular disease (e.g., HDL cholesterol, LDL cholesterol, homocystene and fibrinogen). Development of cardiovascular events is recorded over time including coronary heart disease (and its components; myocardial infarction, coronary death and angina), stroke, intermittent claudication (a peripheral arterial disease), congestive heart failure and cardiovascular disease death. Intense efforts continue to be utilized to gather complete information on every subject. There are some missing data due to subjects moving from the area or discontinuing

participation (which is minimal). The total loss to follow-up is less than 3 percent. The Framingham Heart Study was expanded in 1971 to include a cohort of the offspring of the original participants and their spouses. These data allow for an investigation of the evolution of new detection technologies such as echocardiogram and carotid ultrasound and the study of the effects of genetics on development of cardiovascular and other chronic diseases such as dementia.

3.2. Methodological Issues in Chronic Disease Studies

There are a number of major methodologic issues that arise in longitudinal studies, two are discussed here. The first issue is based on changing definitions of risk factors and outcomes over time. For example, technological advances have resulted in better diagnostic tests for determining the presence or absence of chronic disease. Studies utilizing better diagnostic tests might observe more outcome events and possible different relationships between risk factors and disease. In some chronic diseases (e.g., diabetes) medical specialists have revised the clinical criteria for diagnosing an individual (e.g., different threshold criteria on laboratory tests).

Even the definition of myocardial infarction has changed over time. In the late 1940s, its determination was based mainly on electrocardiogram. Later, enzyme tests, SGOT and CPK, became standard components of the definition of myocardial infarction starting in the mid 1950s and proceeding during the 1960s. In other areas, more sensitive assays have been developed over time for measuring risk factors (e.g., HDL and LDL cholesterol). As modifications occur during a study, analysts must take steps to make the data as comparable over time as possible. The same applies when making comparisons to external studies, these may have employed different definitions and assays.

A second methodological issue in longitudinal studies concerns missing data. Even when intensive surveillance programs are in place, such as those used in the Framingham Heart Study, there are often situations where complete data is not gathered on every subject.

In longitudinal studies of chronic disease, there are instances where data are missing because subjects fail to show up at scheduled examinations, fail to complete certain assessments even when attending the examination, or drop out during the course of the study. These circumstances produce unequal numbers of repeated measurements on different individuals. There are several approaches for performing analysis in the presence of missing data.

First, analysis can be restricted to only those individuals with complete data. This approach is not optimal in terms of efficiency and is biased in some situations. A second approach involves imputing or ascribing values for the missing values and then analyzing the revised dataset. There are sophisticated procedures and software packages available for this imputation and subsequent analysis. This analysis can be biased and can artifically improve precision. A third approach involves analyzing the incomplete dataset (i.e., without attempting to impute values for the missing data).

Statistical techniques and associated computer software (e.g., mixed models) exist that take advantage of all available data and minimize bias that are associated with analysis restricted to only individuals with complete data or analysis of imputed data. These techniques, however, require assumptions about the non-response or the missing data mechanisms. If these assumptions are incorrect, these models can also produce biased results.

The most appropriate analytic techniques in the presence of missing data are those closely tied to the underlying missing data mechanism. When the missingness does not depend on the value of the complete or missing outcome, the data are said to be missing completely at random. Data are missing completely at random if the probability of observing a missing value does not depend on current or future data. For example, if a data monitor forgets to ask a patient if he or she has persistent chest pains (angina) the missingness has nothing to do with this subject's cardiovascular health. A less strict assumption about the missing data mechanism is one in which the missingness is related only to the data observed (and not related to unmeasured or missing data).

This missing data mechanism is called missing at random and the probability of observing a missing value depends on past data but does not depend on current or future data. For example, missing at random results when missingness is related to past cardiovascular health but is independent of unavailable current or future cardiovascular health. Data that are missing completely at random or missing at random are said to be ignorable and to produce a valid analysis it is not necessary to model the missing data mechanism explicitly. Appropriate analysis that include variables related to the mechanism for missingness produce unbiased results.

The final classification of missing data mechanisms is called nonignorable missingness. If the probability of observing a missing value depends on unmeasured current and future data, the missingness is nonignorable. An example would be a subject who fails to show up for an evaluation because his/her health has started to deteriorate. The deterioration continues, and if outcomes were measured, they would reflect the decline. When missing data are nonignorable, it is critical to model the missing data mechansim explicitly in statistical models otherwise results will be biased.

Even with these classifications for missing data and the available statistical techniques and software, there is no formal means to test which mechanism is operating in a given situation. The validity of the analysis often depends heavily upon the assumptions of the technique. Therefore, analysis and interpretation of results in the presence of missing data are often open to criticism. The best recommendation for handling missing data is to avoid it wherever possible.

4. Analytic Techniques for Chronic Disease Modeling

After the sample is selected and the risk factors, the outcomes and the sampling schedule determined, mathematical/statistical modeling is needed to tie these together. Several analytic techniques can be applied to investigate this relation of the risk factors to the development and progression of chronic disease. Some of these are designed specifically to relate baseline risk factors to disease development. Some are able to exploit the time dependent nature of the risk factors and the outcome events. We now describe some popular techniques.

4.1   Logistic   Regression   Analysis: Dichotomous Outcome

Logistic regression analysis can examine and quantify the effects of risk factors on the development of disease. The outcome of interest is dichotomous (e.g., development or non-development of chronic disease over a time period), and the independent variables or risk factors can include continuous or discrete characteristics. The logistic regression model is of the form:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_1 x_1 + \ldots + \beta_p x_p$$

where Y is a dichotomous outcome variable (e.g., 0=no chronic disease, 1=chronic disease) and p=P(Y=1) is the probability of a subject with the disease, x_1, x_2, …, x_p are the risk factors, and β_1, β_2,…β_p are the regression parameters reflecting how the risk factors affect the log of the odds of developing disease. Logistic regression analysis is a very useful technique for analyzing dichotomous outcomes and the individual is considered the unit of analysis.

Logistic regression analysis is appropriate in studies of chronic disease where originally disease free subjects are followed for a pre-specified observation period and at the end of the observation period, each subject can be classified as having developed the disease or not. In many studies of chronic disease, there are often have a number of individuals for whom we do not have data at the end of the observation period and the last time they were observed they had not yet developed disease. Logistic regression can not deal directly with these subjects. The analysts must arbitrarily drop them from analyses or assume a disease status at the end of the observation period. The techniques described in the next section can handle this and other issues that arise in longitudinal studies of chronic disease.

4.2 Survival Analysis: Time to Event Data

Survival analysis includes a set of techniques that deal with time until the event of interest occurs (e.g., onset of disease). It is often the case in studies of chronic disease that there are many patients who do not develop the disease or for whom we do not know if they ever develop the disease. This happens when the disease is rare, when patients are lost to follow-up (e.g., move away but do not develop the disease), when patients die during the observation period but are free of the disease of interest at the time of death, or when they drop out of the study (e.g., due to lack of interest).

In all of these situations, we do not have the time to the development of disease. However, these individuals can contribute a substantial amount of information (up to the end of the observed time period when we know they are disease free) – information which can be utilized through survival analytic techniques. It is this aspect of the data that distinguish survival analysis techniques from other statistical techniques.

These observations in which we know the individual is disease free for some period of time, but do not know if they developed the disease in other time periods are called censored observations. There are several different types of censoring, the most common in studies of chronic disease is right censoring. Right censored obervations are observations in which we do not observe the time to event because if it occurs it occurs after the last observation point.

Some survival models are based on parametric assumptions about the distribution of the survival function, while others are not (parametric and nonparameteric models, respectively). A useful method to characterize survival is by the hazard function (the instantaneous rate of developing disease). There are a number of popular parametric survival models. The exponential model is perhaps the simplest, but assumes constant hazard over time and is therefore not generally applied to chronic disease data. The Weibull distribution model is a generalization of the exponential model and is popular for analyzing chronic disease risk (e.g., cancer

risk) and the hazard function is given by the following:

$$h(t) = \lambda \gamma t^{\gamma - 1}$$

where $\lambda = -\ln(p)/t$ and $p = P$(disease free at time t). The hazard at time t, $h(t)$, increases as t increases for $\gamma > 1$ and decreases as t increases if $0 < \gamma < 1$. The exponential model is a special case of the Weibull model with $\gamma = 1$ (constant risk with time).

Survival analysis methods can be used to assess the effects of risk factors on the development of chronic disease. There are several models that are appropriate for this purpose. A popular parametric model for analysis of chronic disease is the accelerated failure time model whose hazard function is

$$h(t) = e^{\beta'X} h_0(e^{\beta'x} t)$$

where t reflects the time until disease onset, $h_0(t)$ is the baseline hazard at time t (i.e., the hazard if all of the risk factors were set to zero), $\beta'x = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$, x_1, x_2, …, x_p are the risk factors, and β_1, $\beta_2, \ldots \beta_p$ are the regression parameters.

A popular "nonparametric" survival analysis model is the proportional hazards model (also called the Cox regression model), and it is commonly used to assess the relative impact of a set of risk factors measured at a point in time (baseline) on survival and assumes that additive differences in risk factors are related to multiplicative changes in the hazard function.

The proportional hazards model can also be used to assess the impact of time-dependent covariates (i.e., risk factors that change over time) on the hazard function and on survival. This is a particularly useful feature of the model in studies of chronic disease as individuals may undergo procedures during the observation period which alter their prognosis. For example, an individual's risk of cardiovascular disease may change after undergoing coronary artery bypass surgery. The form of the Cox model is:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

where h(t) is the hazard at time t, $h_0(t)$ is the baseline hazard at time t (i.e., the hazard if all of the risk factors were set to zero), and as above x_1, x_2, …, x_p are the risk factors, β_1, $\beta_2,...\beta_p$ are the regression parameters reflecting how the risk factors affect the hazard. The risk factors, x_i above, can be variables measured at some baseline period or variables that vary over time (called time dependent variables).The proportional hazards model is actually a semi-parametric model because the distribution of the underlying hazard is not specified.

Estimating the risk of developing chronic disease per se or assessing the effects of a set of risk factors on the development of chronic disease may be complicated by a common situation in studies of chronic disease, namely, the competing risk of other diseases or death. For example, in studying the relation of risk factors to the development of coronary heart disease the competing risk of someone developing stroke needs to be considered. Similarly, in examining the relation of cigarette smoking to lung cancer the competing risk of developing a heart attack before the lung cancer is a real possibility.

Recently, there have been major efforts to estimate the lifetime risk of developing chronic diseases such as breast cancer, coronary heart disease and Alzheimer's disease. A major methodological issue involves the handling of death which can occur before the chronic disease, such as Alzheimer's disease, develops.

4.3 Longitudinal Data Analysis: Mixed Models, Generalized Linear Models and Generalized Estimating Estimating Equations

A key feature of chronic disease data is the repeated aspect of the measurements. In longitudinal studies with multiple measurements taken on a set of individuals over time, analytic techniques must take into account the correlation between measurements taken on the same individual. An added complexity is the unbalanced nature of the data due to different numbers of measurements taken on different subjects. We now describe some popular methods for analyzing incomplete longitudinal data; mixed models and generalized estimating equations.

Mixed models procedures assume that measurements taken over time are correlated and that regression coefficients vary randomly across subjects according to a specified distribution. In these applications, some of the effects are modeled as fixed (e.g., the effects of risk factors on outcome, called within subjects effects) and some as random (between subject effects). These mixed effects models are also referred to as random coefficients models, growth curve models or hierarchical models. They can also be extended to incorporate time-dependent covariates.

In these mixed effects models a parametric structure is assumed also for the covariances of the repeated measurements. There are many distinct structures that can be assumed, including the independence structure (all observations are independent), compound symmetry (the correlation between any two observations is equal to some common value), autoregressive, and unstructured (no specification of the structure of the correlations).

Currently available statistical computing packages offer many of these structures as options in their mixed models applications. Estimates of the fixed effects and the covariances of the random effects can be estimated using maximum likelihood using Newton-Raphson techniques or the Expectation Maximization (EM) algorithm. The estimates of the covariances are biased because they do not take into account the estimation of the fixed effects and therefore it is recommended that these be estimated using restricted maximum likelihood which produces unbiased estimates. Estimates of the standard errors of effects are robust for large samples.

Mixed models are appealing models for longitudinal data as they are flexible and handle unbalanced data in a highly efficient manner. It is important to note that these models produce consistent estimates (unbiased for large samples) only when data are missing

at random or missing completely at random. These models require careful specification of the fixed and random effects and a covariance structure. When appropriate specifications are made, the final estimates of the fixed and random effects, as well as the magnitude of the variance components are statistically correct and highly informative.

A generalized linear model is a model in which a specific link function (e.g., binomial, Poisson, Gamma) is specified to relate the mean (or expected) value of the outcome to a linear function of the risk factors. This has the effect of transforming the data to a linear model, but involves correct specification of the link or distribution of the outcome variable. Parameters of the model are estimated through maximum likelihood. The appropriateness of the estimates in a generalized linear model are highly dependent on the distributional assumptions.

Generalized estimating equations (GEE) are used to analyze correlated data (e.g., data measured on the same subject over time) that could otherwise be analyzed using a generalized linear model but require fewer distributional assumptions than generalized linear models, making them more appealing. The method of estimation is an extension of least squares.

Generalized estimating equations produce consistent estimates (unbiased for large samples) and robust standard errors for large samples. Generalized estimating equations are appropriate when interest lies in "marginal" effects (i.e., effects averaged over all individuals) rather than subject-specific effects. The approach is now available in many statistical computing packages and again requires specification of a covariance structure. It is appropriate under the asusmption of data missing completely at random.

4.4 Tree-Based Classification Methods

Still another set of techniques for relating risk factors to development of chronic disease are tree-based classification methods. These include a number of applications which are intuitively appealing, many of which are based on a technique called binary recursive partitioning.
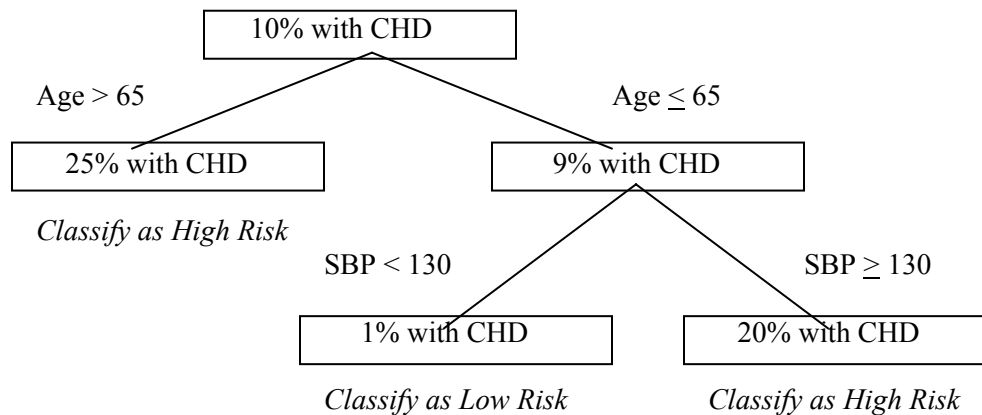
In binary recursive partitioning, a dataset is partitioned first into two distinct groups on the basis of the risk factor that best discriminates the groups in terms of disease status (present or absent). The process is recursive in that this partitioning continues until pre-specified stopping criteria are met (e.g., the final groups represent the last statistically significant splits). The outcome of these analyses is in the form of a clinical prediction rule or algorithm that resembles a tree where the branches represent splits on a risk factor.

Figure 1 illustrates a simple tree where there are two splits. The first split is on the basis of age (over 65 years versus 65 years and younger). A second split is made among those 65 years of age and younger on the basis of systolic blood pressure (less than 130 mm Hg versus 130 or more mm Hg). Persons over 65 years of age have a 25% probability of developing coronary heart disease. Persons 65 years of age and younger with systolic blood pressure less than 130 have a 1% probability of developing CHD, while persons 65 years of age and younger with systolic blood pressure of 130 or more have a 20% probability of developing CHD.

When the outcome is dichotomous (presence or absence of chronic disease) the rule can be used to classify patients, on the basis of specific criteria, as likely or unlikely to develop the disease. The criteria are based on specific values of risk factors. These models are particularly appealing to clinicians as they mirror common practice. For example, a physician might gather information from a patient on his/her risk factors (e.g., systolic blood pressure, smoking status, alcohol consumption), and may conduct a series of laboratory tests (e.g., total Cholesterol level, HDL cholesterol, triglycerides). Based on this information, the clinician can appeal to the empirical tree-based prediction rule to classify the subject as likely or not likely to develop the disease. These methods can also be used to estimate the probability that this patient will develop chronic disease.

Figure 1.   Tree-Based Classification Methods: Example of A Simple
                 Classification Tree for Coronary Heart Disease (CHD)

```
                        ┌─────────────────┐
                        │  10% with CHD   │
                        └─────────────────┘
        Age > 65          ╱             ╲          Age ≤ 65
             ┌──────────────────┐   ┌──────────────────┐
             │  25% with CHD    │   │   9% with CHD    │
             └──────────────────┘   └──────────────────┘
             Classify as High Risk     ╱          ╲
                           SBP < 130  ╱            ╲  SBP ≥ 130
                    ┌──────────────┐     ┌──────────────┐
                    │ 1% with CHD  │     │ 20% with CHD │
                    └──────────────┘     └──────────────┘
                   Classify as Low Risk   Classify as High Risk
```

4.5 Neural Networks

Neural network models are a large class of elaborate mathematical techniques used for developing prediction rules. They are now becoming popular methods for predicting chronic disease. They are very flexible prediction models that can accommodate large datasets (i.e., many risk factors and large sample sizes) and more complex relationships among the variables.

4.6 Model Building

All of the above methods often involve a development phase and a validation phase. Investigators split a dataset into two distinct parts, one part is used for developing the model and the other part is used to evaluate how the model performs (the validation phase).

5. Future Directions

The collection and analyses of chronic disease data have evolved over time to a new level of sophistication. The development of new statistical methodologies for longitudinal data analysis and analysis of complex systems, coupled with advances in statistical computing, have greatly influenced the statistical analysis of chronic disease data. As health care delivery systems continue to strive for quality, more data will be collected and available for analysis of chronic disease (and also for acute and epidemic disease). Longitudinal data will be available on many subjects thereby allowing for more complete investigations of risk factors and interactions between risk factors.

Advances in statistical computing software will also allow for the estimation of more complex statistical models, not restricted to those which assume linear associations between risk factors and chronic disease. Finally, as more data become available on families, analysis of chronic disease will include exploration of genetic factors on the development and progression of disease.

References

Anderson KM, Wilson PW, Odell PM, Kannel WB (1991). An updated coronary risk profile. A statement for health professionals. *Circulation 83*, 356-362.

Beiser A, D'Agostino RB Sr., Seshadri S, Sullivan LM, Wolf PA (2000). Computing estimates of incidence, including lifetime risk: Alzheimer's disease in the Framingham Study. The Practical Incidence Estimators (PIE) macro. *Statistics in Medicine, 19*, 1495-1522.

Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, Lee PW et al. (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet 350*, 469-472.

Concato J, Feinstein AR, Holford TR (1993). The risk of determining risk with multivariable models. *Annals of Internal Medicine 118*, 201-210.

D'Agostino RB Sr., Belanger AJ, Markson EW, Kelly-Hayes M, Wolf PA (1995). Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model. *Statistics in Medicine 14*, 1757-1770.

D'Agostino RB Sr., Griffith JL, Schmid CH, Terrin N (1998). Measures for evaluating model performance. In proceedings of the Biometrics Section American Statistical Association. Biometrics Section 253-258.

Harrell FE Jr., Lee KL, Mark DB (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*, 361-387.

Hosmer DW Jr., Lemeshow S (1989). *Applied Logistic Regression.* New York: Wiley.

Knuiman MW, Vu HT, Segal MR (1997). An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *Journal of Cardiovascular Risk, 4*, 127-134.

Knuiman MW, Vu HT (1997). Prediction of coronary heart disease mortality in Busselton, Western Australia: An evaluation of the Framingham, national health epidemiologic follow-up study, and WHO ERICA risk scores. *Journal of Epidemiology and Community Health 51*, 515-519.

Laird NM. (1988). Missing Data in Longitudinal Studies. *Statistics in Medicine 7*, 305-315.

Lapuerta P, Azen PS, LaBree L (1995). Use of neural networks in predicting the risk of coronary artery disease. *Computational Biomedical Research, 28*, 38-52.

Lloyd-Jones DM, Larson MG, Beiser A, Levy D (1999). Lifetime risk of developing coronary heart disease. *The Lancet, 353*, 89-92.

Long WJ, Griffith JL, Selker HP D'Agostino RB Sr. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computational Biomedical Research, 26*, 74-97.

Segal MR, Bloch DA (1989). A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine, 8*, 539-550.

Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB Sr. (1995). A comparison of performance of mathematical predictive models for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. *Journal of Investigational Medicine, 43*, 468-476.

Seshadri S, Wolf PA, Beiser A, Au R, McNulty K, White R, D'Agostino RB Sr. (1997). Lifetime risk of dementia and Alzheimer's disease: The impact of mortality on risk estimates in the Framingham Study. *Neurology*, *49*, 1498-1504.

Zeger SL, Liang KY and Albert PS (1988). Models for Longitudinal Data: A generalized Estimating Equation Approach. *Biometics, No. 44*, 1049-1060.

Zhang H, Crowley J, Sox HC, Olshen RA (1997). Tree-Structured Statistical Methods. *Encyclopedia of Statistics*.